Gene tree dependence from finite recombination affects summary methods

Wanting He[1], Celine Scornavacca[2], and Yao-ban Chan[1]

https://doi.org/10.5802/fake.doi

Abstract

When inferring the evolutionary history of species and the genes they contain, the phylogenetic trees of genes can be different from those of the species and to each other, due to a variety of causes, including incomplete lineage sorting. We often wish to infer the species tree, but only reconstruct the gene trees from sequences. We then combine the gene trees to produce a species tree; methods to do this are known as summary methods, of which ASTRAL is currently among the most popular. ASTRAL has been shown to be practically accurate in many scenarios through extensive simulations. However, these simulations generally assume that the input gene trees are independent of each other. This is known to be unrealistic, as genes that are close to each other on the chromosome (or are co-evolving) have dependent phylogenies, due to the absence of unlimited recombination between the genes.

In this paper, we develop a model for generating dependent gene trees within a species tree, based on the coalescent with recombination. We then use these trees as input to ASTRAL to reassess its accuracy for dependent gene trees. Our results show that ASTRAL performs more poorly with greater dependence, both when gene trees are known and estimated from sequences. Indeed, the effect of dependence between gene trees is comparable to (if not greater than) the effect of gene tree estimation error. We then re-analyse a 37-taxon mammalian data set; under a realistic recombination rate, the estimated accuracy of ASTRAL decreases substantially (the normalised Robinson-Foulds distance increases by a factor of 4.9) relative to the accuracy previously estimated with independent gene trees, and the effective sample size for this dataset is about one-fifth of the actual sample size. This shows that the impact of gene tree dependence on the accuracy of ASTRAL (and other summary methods) can be significant.

Keywords: Incomplete Lineage Sorting, Species Tree Inference, ASTRAL, Gene Tree Dependence, Gene Tree Estimation Error, Recombination

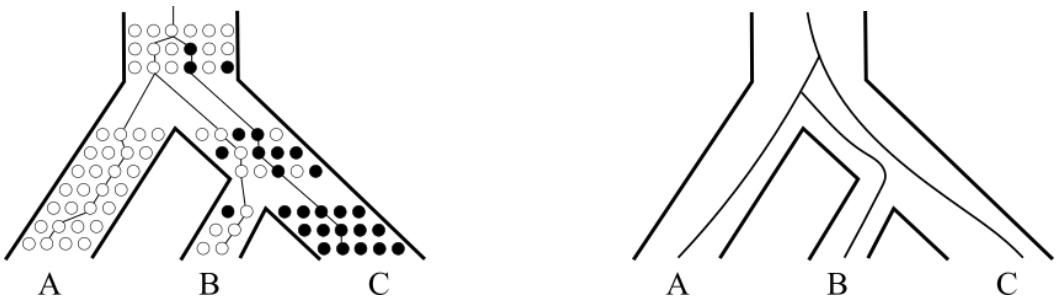
[1]School of Mathematics and Statistics/Melbourne Integrative Genomics, The University of Melbourne, Melbourne, Victoria, Australia, [2]Institut des Sciences de l'Évolution Montpellier, Université Montpellier, Montpellier, France

Correspondence
yaoban@unimelb.edu.au

Introduction

Speciation is an evolutionary process where populations evolve and become distinct species. A species tree, or phylogeny, depicts the history of speciation where leaves represent extant species, internal nodes represent speciation events, and evolutionary distances are represented by branch lengths. Likewise, gene trees depict the evolutionary history of gene families within the genomes of these species. Species trees and gene trees play a vital role in the study of gene and genome evolution, and their reconstruction can give us insight into the history of life on Earth.

When species lineages diverge through speciation, gene copies within these species also diverge. Therefore, gene trees can be thought of as evolving within the branches of the species tree. However, in addition to speciation, various gene-only evolutionary processes can cause gene trees to be distinct from the species tree; these include gene duplication and loss, horizontal gene transfer (HGT), and incomplete lineage sorting (ILS) (Maddison, 1997). In particular, ILS, where multiple gene lineages do not coalesce over a series of speciations, is a major cause of discordance between gene and species trees (see Figure 1 for an example). The standard statistical model for gene trees under ILS is the multispecies coalescent (MSC) model (Pamilo and Nei, 1988; Rannala and Yang, 2003). In this model, each species branch represents a separate population for which Kingman's basic coalescent (Kingman, 1982) is run in a bottom-up fashion, and lineages at the top of each branch are used as input to their parent branches.



(a) Population-level view **(b)** Gene tree view

Figure 1 – An example of incomplete lineage sorting. Circles represent individuals of a population, and each row corresponds to a generation. Originally, the population contains only individuals with an ancestral allele (open circles), and then a mutation introduces a derived allele (filled circles), creating genetic polymorphism. The derived allele does not descend in species A, leaving only individuals with the ancestral allele. The polymorphism persists in the ancestor of species B and C. Subsequently species B retains only the ancestral allele, while species C retains only the derived allele. For this gene family, individuals in population B are more closely related to individuals in population A than C, leading to topological discordance between the gene and species trees.

Because gene trees can be different from the species tree, we need to utilise multiple gene families to infer the species tree. In the past, concatenation has been used to infer the species tree, where sequences from multiple genes are concatenated to form a supermatrix, and then a species tree is obtained by applying traditional phylogenetic inference methods, such as distance-based or maximum likelihood. This paradigm implicitly assumes that all genes are in accord with a single gene tree, which (for the reasons given above) is known not to be true.

More recently, summary methods have been developed, where sequences from different loci are analysed separately to build gene trees, which are then summarised in some way to

generate a species tree. Some summary methods only use gene tree topologies, for example, MP-EST (Liu et al., 2010), NJst (Liu and Yu, 2011), ASTRID (Vachaspati and Warnow, 2015), DISTIQUE (Sayyari and Mirarab, 2016), ASTRAL (Mirarab et al., 2014; Mirarab and Warnow, 2015; Zhang et al., 2018) and STAR (Liu et al., 2009), and some use both gene tree topologies and branch lengths, such as GLASS (Mossel and Roch, 2008) and STEAC (Liu et al., 2009). The input of these methods are mainly rooted gene trees, but some (such as ASTRAL, NJst, ASTRID, and DISTIQUE) can take unrooted gene trees as input. A number of other paradigms are also available, including full-likelihood, Bayesian, and co-estimation methods.

ASTRAL is a popular summary method due to its high accuracy (Ballesteros and Sharma, 2019; Giarla and Esselstyn, 2015) and scalability (Mirarab and Warnow, 2015; Yin et al., 2019). It has been shown to give a statistically consistent estimator of the species tree if input gene trees are sampled under the MSC model (Mirarab et al., 2014), bounded HGT models (Davidson et al., 2015), the general DLCoal model (Markin and Eulenstein, 2021), or the GDL model (Legried et al., 2021; Yan et al., 2022). Furthermore, extensive simulations have been performed studying its accuracy under practical conditions, showing that ASTRAL is highly accurate under the MSC model (Mirarab et al., 2014; Mirarab and Warnow, 2015; Sayyari and Mirarab, 2016; Vachaspati and Warnow, 2015), the presence of both HGT and ILS (Davidson et al., 2015), and with gene filtering when ILS is low to moderate (Molloy and Warnow, 2018). An important practical consideration is whether the gene trees are known exactly or are estimated from sequences with some potential error. It is well known that the accuracy of all summary methods, including ASTRAL, is affected by gene tree estimation error (DeGiorgio and Degnan, 2014; Huang and Knowles, 2016; Lanier and Knowles, 2015; Molloy and Warnow, 2018; Patel et al., 2013).

Although the accuracy of ASTRAL has been extensively studied by simulations, almost all simulations were performed with gene trees that were simulated independently from the model of choice. In effect, this assumes no recombination within a gene and unlimited recombination between genes, both of which are known not to be realistic. The former assumption has been tested (Lanier and Knowles, 2012; Zhu et al., 2022) and found to have relatively little practical effect on the accuracy of species tree inference.

In this paper, we focus on the latter assumption; in reality, gene trees are dependent because genes are located near to each other on the same chromosome, which causes them to have related histories. Indeed, in the absence of recombination, adjacent genes will have identical phylogenies. Recombination breaks this dependence when and where it occurs, but because it occurs at a finite rate, the gene trees of nearby genes are not fully independent. Other causes, such as functional gene linkage, may also induce dependence between gene trees (Barker and Pagel, 2005). Thus, by assuming that the input gene trees are independent, previous simulations may have mis-estimated the accuracy of ASTRAL.

This effect was previously studied by Wang and Liu (2016), who found a significant impact on the accuracy of ASTRAL when gene trees are dependent. However, their methodology attempted to delineate loci from contiguous genomes using inferred or known breakpoints; this has the effect of very strong dependence between neighbouring gene trees, which may not always occur. Additionally, Conry (2020) found that recombination between exons has little effect on the accuracy of species tree inference, but from simulations only on four-taxon species trees; the effect may become more noticeable with larger trees.

The effect of recombination on the evolutionary history of a genome is well-known in population genetics, where the standard model for a single population is the coalescent with recombination (Hudson, 1983). This produces an evolutionary history that contains both coalescent and recombination events, known as the ancestral recombination graph (ARG) (Griffiths and Marjoram, 1997). The program `ms` (Hudson, 2002) was an early tool for simulating the ARG; more recently, `msprime` (Baumdicker et al., 2022; Kelleher et al., 2016) was developed as a high-speed, large-scale successor, offering both rapid simulation and efficient data storage. The problem of inferring the ARG from genome sequences has been extensively studied, with great progress in recent years (Kelleher et al., 2019; Mahmoudi et al., 2022; Rasmussen et al., 2014; Speidel et al., 2019). Earlier methods, such as ARGWeaver (Rasmussen et al., 2014), achieved efficiency by using an approximation to the coalescent with recombination known as the sequentially Markov coalescent (SMC) (McVean and Cardin, 2005), which assumes that gene trees are Markovian along the genome.

Our aim in this paper is to reassess the accuracy of ASTRAL under conditions when gene trees are dependent. In addition, we also explore other important factors that may affect the accuracy of species tree inference, including the number of loci, the amount of ILS, recombination rate, and the dependence structure. We generalise the two-locus, 3-taxon model of Slatkin and Pollack (2006), and derive a probabilistic method for generating dependent gene trees for multiple species and loci. We then use the generated trees as input to ASTRAL to estimate the species tree. When re-analysing a 37-taxon mammalian dataset (Mirarab et al., 2014), we find that under realistic recombination rates, the accuracy of ASTRAL may have been significantly over-estimated due to the assumption of infinite recombination between loci, with an effective sample size roughly one-fifth that of the actual sample size. This indicates that gene tree dependence may be a significant factor when estimating the accuracy of species tree inference, potentially even more so than gene tree estimation error.

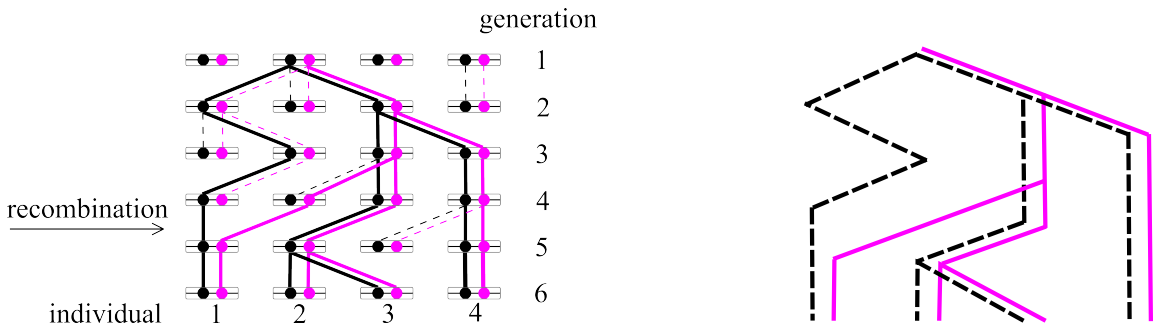
Methods

We first describe a method to generate dependent gene trees, then use these trees as input to ASTRAL in a simulation study. This allows us to determine the accuracy of ASTRAL under conditions of varying gene tree dependence.

Generating dependent gene trees

To generate dependent gene trees within a species tree, we extend the model of Slatkin and Pollack (2006), which considered two loci of haploid individuals in a single Wright-Fisher (panmictic) population. In that paper, they devised a probabilistic model of two linked loci in three species based on the coalescent with recombination, using it to calculate the probabilities that the gene trees in the loci are concordant with the species tree and with each other. Here, we extend the model in a natural way to multiple species and loci, but more importantly present an algorithm to simulate a gene tree in one locus conditional on a known tree in a linked locus. This idea was also used in Li et al. (2021) to model so-called 'linked duplications'. Figure 2 shows an example of our model (described below).

In the absence of recombination, the parents of all individuals are the same between the loci, and so the genealogies will be the same. We model this by considering the known genealogy in the first locus as a 'guide tree', to which lineages in the second locus can 'coalesce', representing



(a) Population view. A recombination event happens in the leftmost individual of generation 5, which leads to a divergence of lineages between the two loci, and thus to topological discordance between the two gene trees.

(b) Gene view. The black tree can be considered as a 'guide tree' which the pink tree can coalesce and uncoalesce with.

Figure 2 – An example of the two-locus model. The black and pink lines represent the genealogies in the first and second loci respectively.

that the lineages belong to the same individual. If a lineage is coalesced with the guide tree, it must then follow the ancestry of that lineage if no recombination occurs. Because we observe extant samples of multiple loci in the same individuals, the lineages in the second locus begin coalesced with the guide tree at the present time. Thus, in the absence of recombination, the two trees will be identical.

Recombinations occur between the two loci at a rate of R per individual per coalescent unit. (If there is a genomic distance of d sites between loci, then in population genetics notation, $R = 2N_e rd$, where N_e is the effective population size and r is the recombination rate per site per generation.) When a recombination occurs, the two loci in an individual have distinct parents in the preceding generation. We represent this as the lineage in the second locus 'uncoalescing' from the guide tree, so that it no longer follows the ancestry of the first locus. However, recombination is an event that happens to a single individual/lineage, so other lineages are unaffected. Note in particular that coalesced lineages in the second locus cannot uncoalesce from each other, as this does not represent recombination.

Proceeding backwards in time, uncoalesced lineages in the second locus can either coalesce with a lineage of the guide tree (whether or not it is already coalesced with a lineage in the second locus) or with each other. These coalescences occur at rates consistent with the multispecies coalescent. Thus if there are k_u uncoalesced lineages and k_g guide tree lineages in a population at a particular time, the uncoalesced lineages will coalesce with each other at a rate of $\binom{k_u}{2}$ and with a guide tree lineage at a rate of $k_u k_g$ (in coalescent units). The times of coalescences between the guide tree lineages are specified by the guide tree. As with the basic coalescent, this process continues until all lineages in the second locus have reached their most recent common ancestor (MRCA), which is not necessarily the MRCA in the first locus.

We extend this process to multiple species in a standard way. To generate multiple dependent gene trees, we first generate a gene tree under the MSC. Each gene tree is then used as a guide tree for the subsequent gene tree. This is essentially the sequentially Markov coalescent (SMC) (McVean and Cardin, 2005), sampled at constant intervals along the genome.

We note that there already exist several tools for simulating gene trees under the full coalescent with recombination, such as `ms` and `msprime`. We have elected not to use these tools, as `ms` is too slow, while `msprime` does not currently accept non-ultrametric species trees, which we

have. Although our model lacks the long-range dependence structure of the full coalescent with recombination, it is faster to simulate and retains more flexibility; for example, it can be used in the case of a non-linear dependence structure between genes (for example, when dependence indicates functional rather than proximity relationships). Nevertheless, we conduct some tests on an ultrametric tree to determine that the results from our model do not differ significantly from those produced using gene trees generated by `msprime`.

The described model assumes a constant amount of recombination between loci, which is not biologically realistic due to varying gene positions and recombination rates along the genome. We also explore an alternative way of generating dependent gene trees where we generate independent 'blocks' of gene trees, where block boundaries are determined randomly along the genome. Within each block, we generate trees with a constant recombination rate R as above.

We perform this scenario in two different ways. First, we keep the recombination rate R fixed and vary the number of independent blocks, while keeping the total number of gene trees fixed. This also varies the 'overall' recombination rate, so to observe the effects for a fixed overall recombination rate, we have a second scheme in which we vary the recombination rate within each block while fixing the average value of $\frac{R}{1+R}$ between trees. An intuitive explanation for this expression is that lineages coalesce at a rate of 1 and uncoalesce with a rate of R; therefore $\frac{R}{1+R}$ is approximately the fraction of time that a new tree spends uncoalesced from the guide tree, and hence quantifies the amount of new information each tree adds to the sample.

Simulations

We explore the performance of ASTRAL on a 37-taxon mammalian species tree, which has previously been used to study the accuracy of ASTRAL for independent gene trees (Mirarab et al., 2014). The species tree, shown in Figure 3, was previously estimated with MP-EST (Liu et al., 2010) on the biological dataset from Song et al. (2012), containing 447 genes with average length 3099 bp. We re-estimate the branch lengths as specified in Supporting information, while keeping the same topology.

In order to reassess the accuracy of ASTRAL with dependent gene trees, we perform simulations with the same parameter settings as in Mirarab et al. (2014), adding the dependency structure: given a species tree, we generate a set of N dependent gene trees using our model above, varying R. We then use these trees as input to ASTRAL to infer a species tree, and evaluate the accuracy using the normalised Robinson-Foulds (RF) distance (Robinson and Foulds, 1981). The RF distance measures the difference between the true and inferred species trees by summing the number of clades present in one tree but not the other.

Additionally, we rescale the simulated dependent gene trees to produce average branch lengths equal to those estimated from the dataset (Song et al., 2012), then evolve gene sequences along the trees with `Pyvolve` (Spielman and Wilke, 2015), under the GTR+Γ model of site evolution (Tavaré, 1986) with parameters estimated from the biological dataset. As in Mirarab et al. (2014), we have three sequence lengths: 500 bp, 1000 bp, and a mixture where half of the sequences are 500 bp and the other half are 1000 bp. We then use IQ-TREE (Minh et al., 2020; Nguyen et al., 2015) to estimate gene trees from these sequences, and use the estimated gene trees as input to ASTRAL. This allows us to study the effect of gene tree estimation error.

From initial results, we found that accuracy varies the most when the recombination rate R lies between 0 and 1, so we use this range for our recombination rate parameter. When R is equal

to 0, all gene trees are identical; as it increases, there is less dependence between neighbouring gene trees, and when it is infinite, the gene trees are independent.

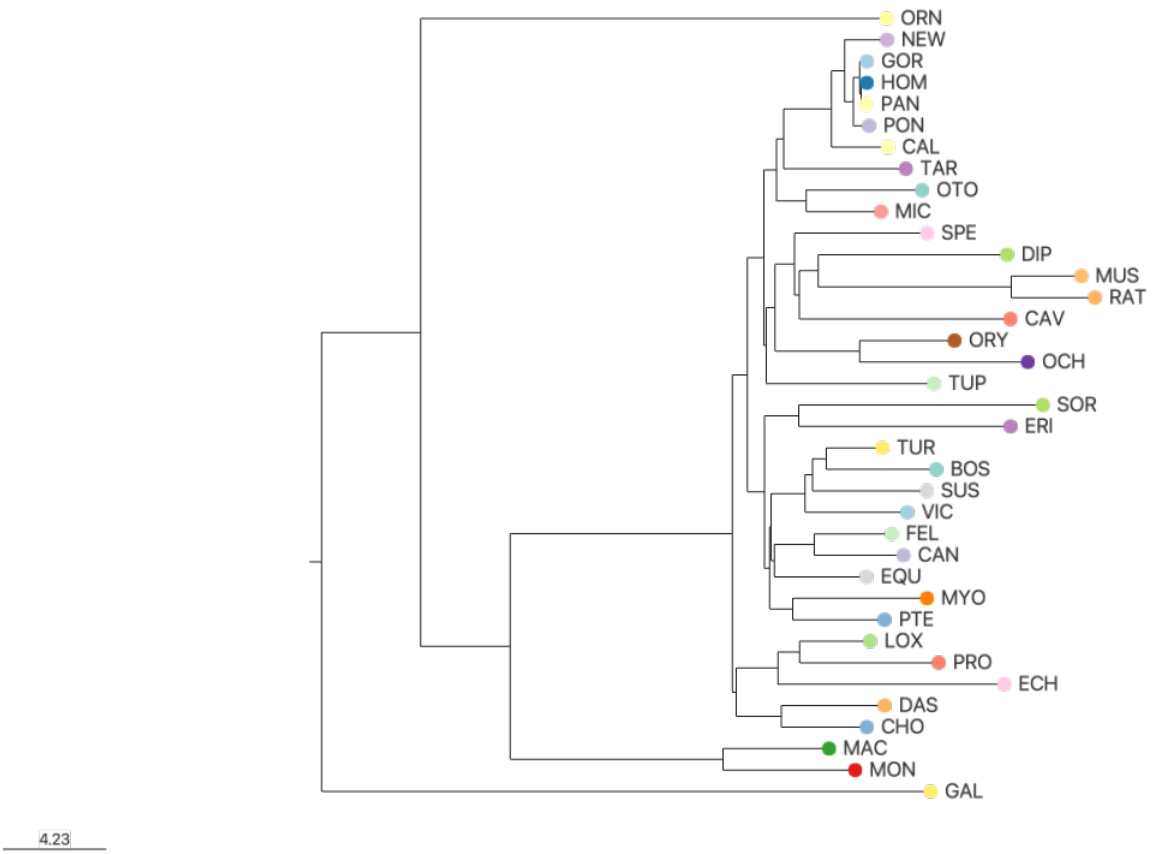


Figure 3 – The 37-taxon mammalian species tree (see species names in Table 15 in Supporting information).

It has been shown that ASTRAL performs worse with an increased amount of ILS (Mirarab et al., 2014), so we are interested in the performance of ASTRAL with dependent gene trees in this case. We multiply the branch lengths of the species tree by 0.2 (denoted by $0.2\times$), which is equivalent to multiplying the effective population size by 5, and repeated the simulation above. Finally, we also fix the number of gene trees N to be 200, and evaluate the accuracy of ASTRAL with different amounts of ILS, with branch length multipliers ranging from $0.2\times$ to $5\times$ as in Mirarab et al. (2014).

The parameter settings for the simulations are summarised in Table 1. We perform 100 replicates for each parameter setting for true gene trees, and 20 replicates for estimated gene trees (as gene tree estimation is computationally intensive). In addition, we only use $N \leq 800$ gene trees per dataset for estimated gene trees. The same simulation settings are used when varying recombination rate along the genome, except that we use 100 replicates and fix the overall recombination rate (as described above) to be a realistic value estimated from the biological dataset.

To verify that the our results are not biased by the specific species tree used, we repeat the above simulations (for true gene trees only) on a tree of 16 fungi (Butler et al., 2009; Legried et al., 2021; Markin and Eulenstein, 2021; Rasmussen and Kellis, 2012; Wu et al., 2014), shown

Table 1 – Simulation parameters.

Simulation	R	N	ILS[1]
Default simulations	[0, 1]	[100, 3200]	$1\times$ branch lengths
Increased amount of ILS	[0, 1]	[100, 1600]	$0.2\times$ branch lengths
Different amounts of ILS	[0, 1]	200	$[0.2, 5]\times$ branch lengths

[1] Multiplying branch lengths of the species tree by α is equivalent to dividing the effective population size by α.

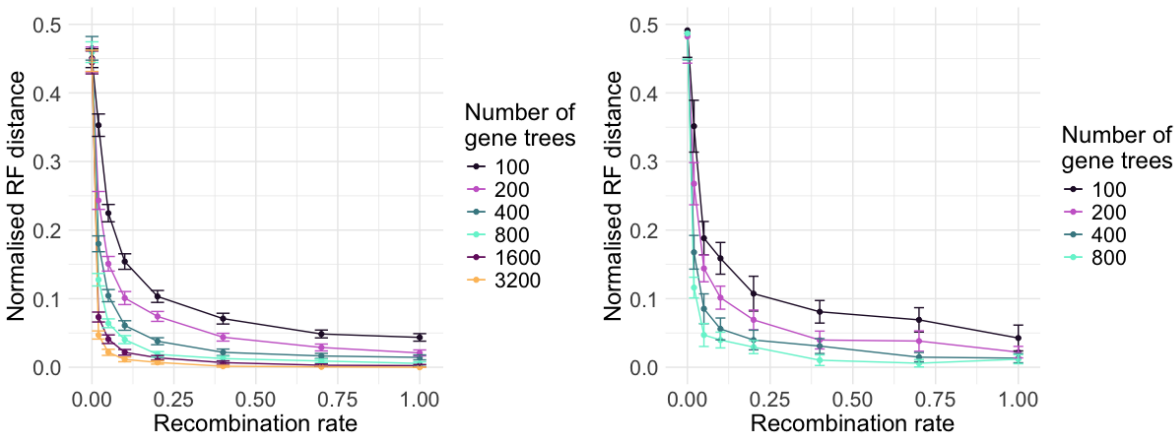
in Figure 19 in Supporting information. This fungal tree is ultrametric and has been studied extensively in simulations aiming to show that the accuracy of ASTRAL is not affected by data with paralogs (Yan et al., 2022). For this dataset, we vary R from 0 to 0.5 following an initial exploration, and we use up to 3200 gene trees for the 'increased ILS' simulation. Otherwise, we use the same simulation settings shown in Table 1.

To compare the performance of our model with `msprime` (Baumdicker et al., 2022; Kelleher et al., 2016), which is restricted to ultrametric species trees, we also conducted simulations under `msprime` on the fungal tree, using the same parameter settings, and compared these results to those generated from our model.

Results

Default simulations

For the first simulation, we study the accuracy of ASTRAL in relation to the recombination rate between gene trees R and the number of input gene trees N. We show the results for both true and estimated gene trees in Figure 4.



(a) True dependent gene trees (100 replicates) **(b)** Estimated dependent gene trees (20 replicates)

Figure 4 – The accuracy of ASTRAL vs recombination rate, averaged over a number of replicates. Error bars represent standard error.

In Figure 4a, we show the results for true gene trees. When $R = 0$, all gene trees must be identical, so the inferred species tree is equal to the single gene tree. Thus the accuracy of ASTRAL is unaffected by the number of input gene trees. As R increases, the normalised RF distance decreases — ASTRAL becomes more accurate with less dependence among gene trees. As $R \to \infty$, the accuracy converges to the results of Mirarab et al. (2014) for independent gene trees.

For constant R, ASTRAL becomes more accurate as the number of gene trees N increases, which is the same as with independent gene trees. It has been shown (Chan et al., 2022) that with independent gene trees, the normalised RF distance decreases exponentially with respect to N. In Figure 21 in Supporting information, we show a log-plot of normalised RF distance against N. It is inconclusive from this plot if the exponential asymptotic behaviour still holds for finite recombination rate; potentially it still does, but may take longer to reach the limiting behaviour.

Figure 4b shows the results for estimated gene trees from mixed sequence lengths, with detailed results for 500 bp and 1000 bp available in Supporting information. The three sequence lengths show minimal differences. In Figure 4b, we see that similar patterns hold when the gene trees are estimated from sequences. Overall, the accuracy is slightly lower for estimated gene trees at all recombination rates, as expected. This indicates that the accuracy of ASTRAL can be adversely affected simultaneously by both gene tree estimation error and gene tree dependence. In Figure 5, we compare the relative effects of these factors with recombination rate fixed to two realistic values (estimated in the Estimating the recombination rate section) of 0.317 and 0.055. We can see that for the two estimated recombination rates, the loss in accuracy from gene tree dependence is considerably larger than the corresponding loss from gene tree estimation error.

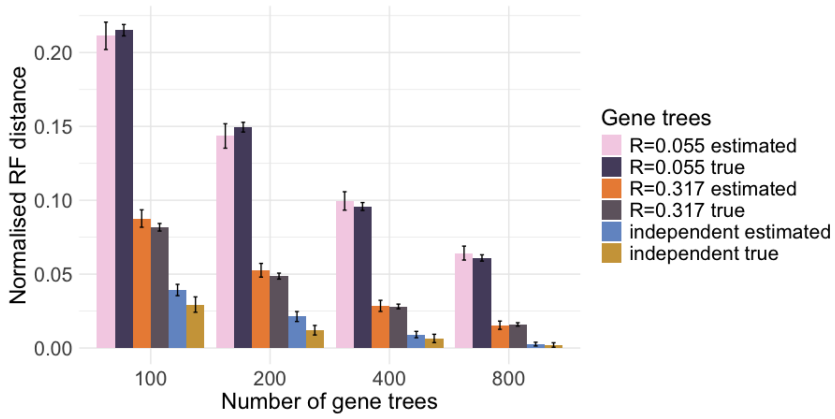


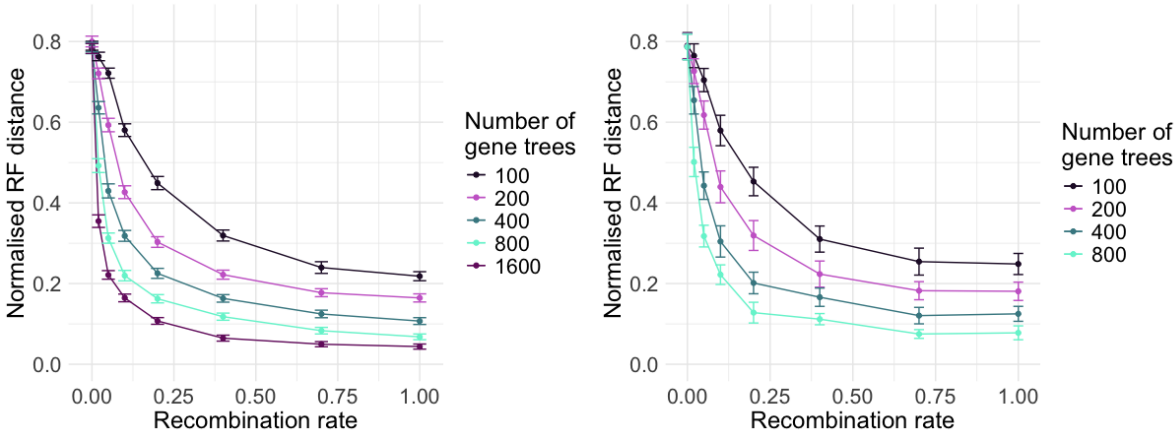
Figure 5 – The accuracy of ASTRAL with and without gene tree estimation error and dependence. The accuracy is shown against the number of gene trees for $\mathbf{R = 0.317}$ and $\mathbf{R = 0.055}$. Error bars represent standard error.

Moreover, the impact of gene tree estimation error appears to be less pronounced for lower recombination rate (more dependence). A possible explanation for this is that in these cases, there is more information available for each tree as multiple gene sequences correspond to the same tree, resulting in less estimation error.

Increased amount of ILS

We repeat our simulations above with an increased amount of ILS ($0.2\times$ branch lengths). We observe similar patterns as in the Default simulations section, as shown in Figure 6, with results for 500 bp and 1000 bp sequence lengths in Supporting information. As before, ASTRAL becomes more accurate with less dependence and with more gene trees. Overall, the accuracy is much poorer for increased ILS, as expected.

In Figure 7, we consider the relative effects of gene tree dependence and estimation error under increased amounts of ILS and two realistic recombination rates $R = 0.317$ and $R = 0.055$.



(a) True dependent gene trees (100 replicates) **(b)** Estimated dependent gene trees (20 replicates)

Figure 6 – The accuracy of ASTRAL vs recombination rate with increased ILS (0.2×). Error bars represent standard error.

We see that the effect of gene tree dependence is even more pronounced relative to the effect of gene tree estimation error when ILS is increased.

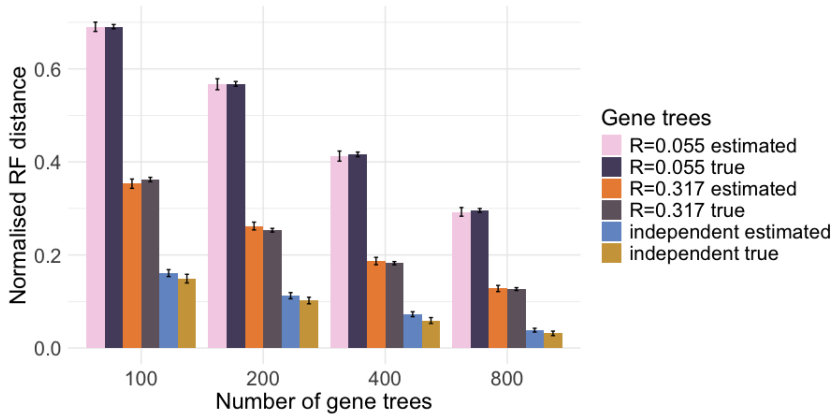
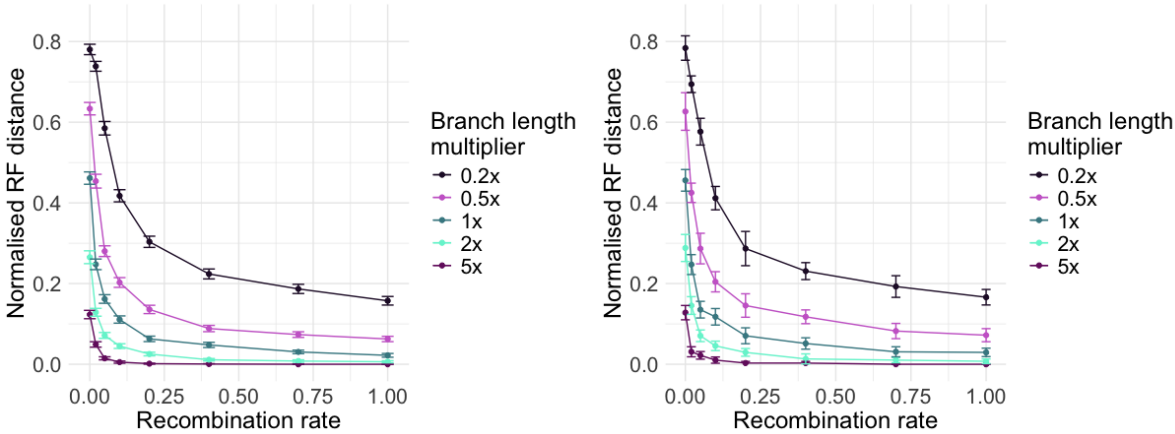


Figure 7 – The accuracy of ASTRAL with and without gene tree estimation error and dependence, with increased ILS (0.2×). The accuracy is shown against the number of gene trees for $\mathbf{R = 0.317}$ and $\mathbf{R = 0.055}$. Error bars represent standard error.

Different amounts of ILS

We also study the effect of different amounts of ILS on the accuracy of ASTRAL when gene trees are dependent. In Figure 8, we fix the number of gene trees to 200 and vary the branch length multiplier. We observe similar patterns of normalised RF distance with respect to recombination rate as above. Overall, ASTRAL performs better with less ILS, as observed for independent gene trees in Mirarab et al. (2014).

In Figure 9, we compare the effects of gene tree dependence and estimation error. As the amount of ILS increases, the effect of gene tree dependence becomes larger relative to gene tree estimation error. Conversely, when ILS decreases, so does the relative effect of gene tree dependence.



(a) True dependent gene trees (100 replicates) **(b)** Estimated dependent gene trees (20 replicates)

Figure 8 – The accuracy of ASTRAL vs recombination rate with different amounts of ILS for N = 200. Error bars represent standard error.

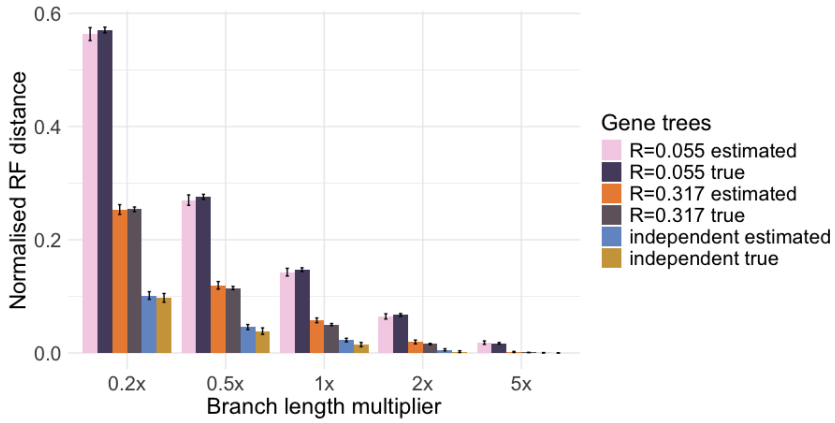


Figure 9 – The accuracy of ASTRAL with and without gene tree estimation error and dependence, with different amounts of ILS. The accuracy is shown against the number of gene trees for **R = 0.317** and **R = 0.055**. Error bars represent standard error.

Decreasing dependence by thinning trees

ASTRAL is constructed under the assumption that the gene trees are independent, and our results above show that it is more accurate in this case. It is common practice to ensure that this is the case by 'thinning' the gene trees, by only considering trees that are a certain distance apart in the genome. If this distance is chosen to be sufficiently large, the resulting trees can be considered to be mostly 'independent' and thus satisfy the assumption of ASTRAL. On the other hand, thinning the trees in this way will reduce their number.

To study this effect, we simulate 1000 dependent gene trees from the 37-taxon mammlian species tree for several recombination rates (0.2, 1, and 5). We then thin our dataset by subsampling every *n*th tree, for *n* ranging from 1 to 50, and use the resulting trees as input into ASTRAL. This is replicated 1000 times, with the results shown in Figure 10.

From this Figure, it is clear that using the full gene tree set (no thinning) is always the most accurate; accuracy decreases monotonically as the subsampling interval increases. Thus, it is always preferable to use a larger number of gene trees, even if they are dependent, over a smaller number of independent gene trees.

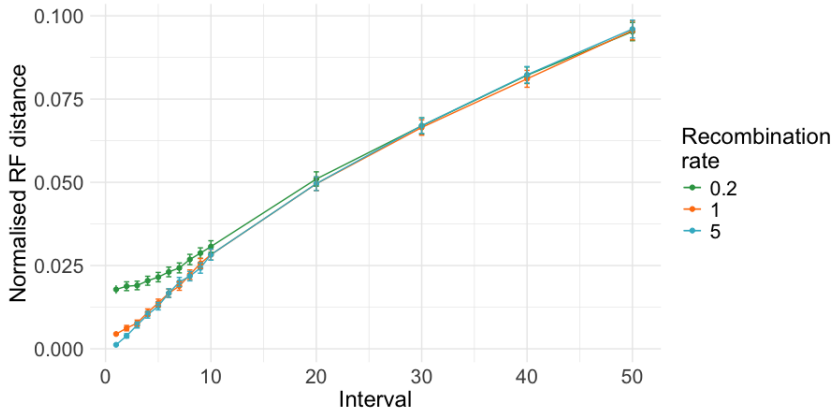


Figure 10 – Accuracy of ASTRAL when thinning dependent gene trees for the 37-taxon mammalian species tree over 1000 replicates. Error bars represent standard error.

Varying recombination along the genome

We observe the effect of varying recombination along the genome by generating independent 'blocks' as described at the end of the Generating dependent gene trees section. We quantify the variability of the recombination rate by the proportion of independent gene trees (i.e., the proportion of gene trees that start a new block, which is the number of blocks divided by N). The results for fixed R (and therefore varying overall recombination rate) are shown in Figure 22 in Supporting information. As expected, we observe that ASTRAL becomes more accurate as the proportion of independent trees (and thus the overall recombination rate) increases.

In Figure 11 (see also Figure 23 in Supporting information), we fix the overall recombination rate to be roughly equivalent to the uniform dependence case with $R = 0.317$ (and $R = 0.055$ in Figure 23). We observe that as the number of blocks increases, the accuracy of ASTRAL degrades slightly; this effect is more noticeable for smaller numbers of gene trees. Thus more dependence within smaller blocks is slightly worse than less dependence within larger blocks. At some point, the recombination rate within blocks becomes 0, and the trees within each block are forced to be identical; this is equivalent to having a smaller number of completely independent trees, and results in the lowest accuracy.

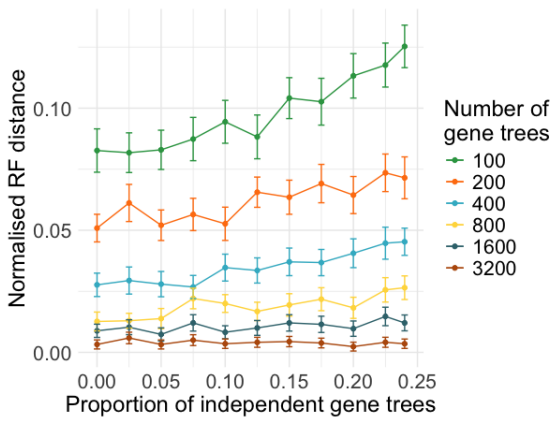


Figure 11 – The accuracy of ASTRAL vs the proportion of independent gene trees. The accuracy is averaged over 100 replicates, with fixed overall recombination rate. Error bars represent standard error.

16-taxon fungal species tree

In Figure 24 in Supporting information, we observe similar patterns to those observed for the 37-taxon mammalian species tree, but with overall higher accuracy. This indicates that our results are not specific to the mammalian species tree.

We also generate dependent gene trees with `msprime`, with results shown in Figure 24. It can be seen that although `msprime` results in a slightly lower accuracy, the results for the two models are in general very similar.

Real data experiments

To study the effect of gene tree dependence on a real biological system, we re-analyse the gene trees from Song et al. (2012) on the 37-taxon mammalian species tree studied above. This dataset contains 424 effective genes (after 21 genes with mislabelled sequences and 2 outliers were removed from the original 447-gene dataset) with average length 3157 bp. We re-estimated the gene trees with IQ-TREE (Minh et al., 2020; Nguyen et al., 2015).

Estimating the recombination rate

The recombination rate between gene trees R is a critical factor in the accuracy of ASTRAL, so we first estimate the recombination rate between trees in this dataset. Although recombination rate between trees can vary in reality due to a number of factors (varying recombination rate per genomic unit, varying genomic distance between genes, or genes being on different chromosomes), we assume for the purposes of this analysis that consecutive genes are adjacent with constant recombination rate between them.

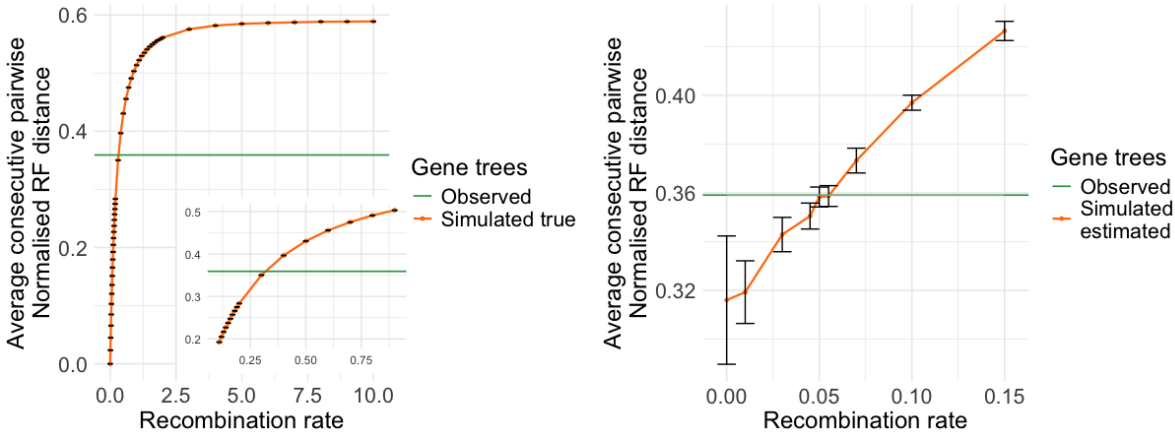
To estimate the recombination rate, we match the average consecutive pairwise normalised RF distance between the real trees to those from simulations. The average consecutive pairwise normalised RF distance of the real gene trees is 0.359 (pairs of genes that are in different chromosomes are not taken into account); this is significantly different from the value for simulated independent gene trees on the same species tree (0.585; a t-test of the difference gives p-value $< 2.2 \times 10^{-16}$). This indicates that the real gene trees are not independent.

We calibrate the recombination rate of our simulations to match this pairwise normalised RF distance. For a range of recombination rates from 0 to 10, we simulate the same number of gene trees (true or estimated), and calculate the average consecutive pairwise normalised RF distance. The results are shown in Figure 12. To estimate the recombination rate, we perform a binary search on the range $[0, 1]$ to reduce sampling variation, and estimate $R = 0.317$ for true gene trees, and $R = 0.055$ for estimated gene trees. The variation between these figures also provides a sense of the accuracy of this procedure.

From our simulations, the accuracy of ASTRAL with 424 independent genes is 0.0047 average normalised RF distance with standard error 0.0004. However, under the recombination rate $R = 0.317$, the accuracy is approximately 0.0267 average normalised RF distance with standard error 0.0008. Thus, by ignoring dependence between gene trees, we are overestimating the accuracy of ASTRAL for this dataset by a factor of 4.9.

Effective sample size

The multiplicative decrease in normalised RF distance due to gene tree dependence varies with the number of gene trees for fixed recombination rate, so it does not provide an overall



(a) True gene trees **(b)** Estimated gene trees

Figure 12 – The average consecutive pairwise normalised RF distance between gene trees. Error bars represent standard error; the error for the observed gene trees is too small to see.

measure of the effect of gene tree dependence. In order to obtain such a measure, we wish to quantify the amount of information or statistical power available in a sample under gene tree dependence; in other words, the effective sample size (*ESS*). To determine the *ESS* for this dataset, we first calculate the average normalised RF distance for ASTRAL for the estimated recombination rate $R = 0.317$. We then use a binary search to determine the number of independent gene trees required to produce the same normalised RF distance. For the real dataset size of 424 gene trees, we determine an effective sample size of 101, roughly one-fifth of the actual sample size.

More generally, we can repeat this procedure for any sample size to fit a relationship between actual and effective sample size for the recombination rates $R = 0.317$ and $R = 0.055$. We show the results in Figure 13; here we see that the relationship between *ESS* and *N* is approximately linear. Fitting a regression line through the origin gives the relationship

$$ESS = 0.21 \times N \text{ for } R = 0.317;$$

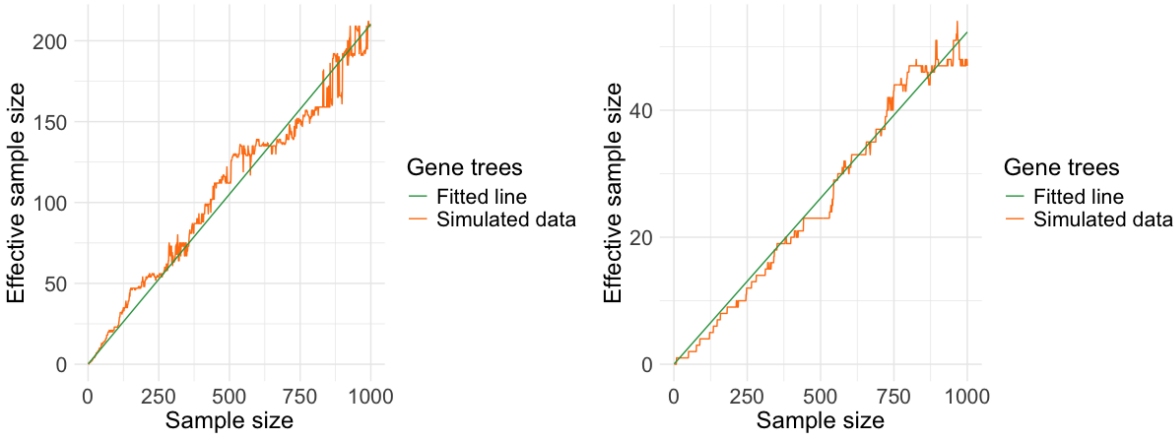
$$ESS = 0.05 \times N \text{ for } R = 0.055.$$

This quantifies the effect of gene tree dependence for this recombination rate, for any sample size.

Comparison with Wang and Liu

Wang and Liu (2016) also studied the effect of gene tree dependence on the accuracy of ASTRAL. To do so, they generated gene trees, and then sequences from the trees, under the (multispecies) coalescent with recombination with `ms` (Hudson, 2002). As they also studied the effect of breakpoint inference on ASTRAL, they compared five cases:

(1) LD1000: Loci of 1000bp were sampled along the sequences at intervals estimated to be (near-)independent from linkage disequilibrium plots; gene trees were estimated from the loci with FastTree (Price et al., 2009, 2010).
(2) LD100: As above, but with locus lengths of 100bp.



(a) R = 0.317 **(b) R = 0.055**

Figure 13 – The effective sample size of ASTRAL under dependence.

(3) IBIG (inferred breakpoints, inferred gene trees): Recombination breakpoints were inferred and used to partition the sequences. Gene trees were then estimated from the partitions with FastTree.

(4) TBIG (true breakpoints, inferred gene trees): As above, but known breakpoints (from `ms`) were used.

(5) TBTG (true breakpoints, true gene trees): True gene trees (from `ms`) were used.

As we do not study the effect of breakpoint inference, a direct comparison can be made only with the last two cases (TBIG and TBTG).

Apart from the model difference (which, as shown above, is not significant), the two major differences between our study and the TBIG/TBTG cases of Wang and Liu (2016) are:

- They use the entire sequences, and partition at every breakpoint; thus their gene trees are separated by exactly one recombination (and therefore highly dependent).
- They do not explicitly control the number of gene trees, and do not directly compare with (the same number of) independent gene trees.

By varying the recombination rate between gene trees, we are able to study scenarios of lesser dependence (such as might occur when we recognise orthologous genes rather than aligning and partitioning whole-genome sequences). By controlling the number of gene trees, we are able to directly compare with independent trees, and thus quantify the effect of gene tree dependence.

To compare our numerical results with theirs, we use our model on the 9-taxon mouse species tree of Wang and Liu (2016), shown in Figure 25 in Supporting information. We calibrate the recombination rate R by calculating the average number of recombinations between consecutive gene trees (from 10000 replicates), and using a binary search so that this average is 1. By repeating this process 20 times, we estimate an equivalent recombination rate of $R = 0.09$ (standard error 0.0001) for this tree.

We also repeated this process for the 37-taxon mammalian species tree above, finding an equivalent recombination rate of $R = 0.0036$ (standard error 0.00001), which is very low compared to our estimated recombination rate of 0.317. This is expected, since the genes in our dataset are not extracted from partitioned sequences, and suggests that in this case the effect of dependence will not be as pronounced as in Wang and Liu (2016).

We then fix the number of gene trees to the average produced by `ms` for the same parameters as Wang and Liu (2016); for the population recombination rates $\rho = 100, 200, 1000$ that they used, this results in 867, 1752, and 8739 gene trees respectively. We then perform simulations as above for gene trees that are dependent and independent. To estimate the effect of gene tree estimation error, we also evolve sequences from true gene trees under the HKY85+Γ (Hasegawa et al., 1985) model with $\alpha = 0.8$, as in Wang and Liu (2016), and then estimate gene trees with FastTree (Price et al., 2009, 2010). Due to computational constraints, we run 100 replicates for inferred gene trees and 1000 replicates for true gene trees.

In Figure 14, we observe again that for these parameters, the effect of gene tree dependence on ASTRAL is far greater than the effect of gene tree estimation error; indeed, for independent trees, the species tree is always inferred without error. This reinforces our conclusion that gene tree dependence is a significant issue, but also indicates that the process followed by Wang and Liu may overestimate the effect of gene tree dependence in a realistic setting.

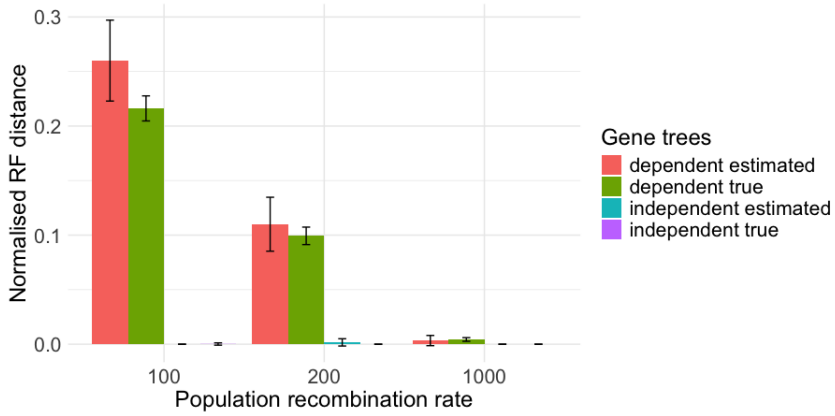


Figure 14 – The accuracy of ASTRAL for the empirical tree of Wang and Liu (2016) for recombination rate R = 0.09. Error bars represent standard error.

We do note that the numerical accuracies we obtain are slightly different from the equivalent cases in Wang and Liu (2016); apart from the differences between the models themselves, we also fix the recombination rate, rather than the number of recombinations between gene trees, and fix the number of gene trees. Calibrating recombination rates by matching the accuracy of ASTRAL between the models results in a larger recombination rate between 0.5 and 1, but with similar results: there is a significant decrease in the accuracy of ASTRAL introduced by gene tree dependence, which dwarfs the smaller effect of gene tree estimation error.

Discussion

In this paper, we developed a model to simulate dependent gene trees within a species tree under a realistic process of incomplete lineage sorting and recombination. We then reassessed the accuracy of ASTRAL for when gene trees are dependent, including both true and estimated gene trees. Our results show that ASTRAL performs more poorly as the dependence between gene trees increases, and gene tree estimation error adversely affects the accuracy of ASTRAL under dependence. In addition, we estimated a realistic recombination rate from biological gene trees, and the estimated effective sample size of ASTRAL. We find that when taking gene tree dependence into account, the accuracy of ASTRAL decreases by a factor of 4.9, and the effective sample size by a factor of 5 for the mammalian dataset. This suggests that simulation studies that

use independent gene trees may be significantly overestimating the accuracy of ASTRAL. Other summary methods that assume independent gene trees may also suffer from the same problem.

In particular, comparing the relative effects of gene tree dependence and gene tree estimation error for a realistic recombination rate suggests that the former has a greater effect on the accuracy of species tree inference. As gene tree estimation error is a well-known issue in phylogenetic inference, this suggests that gene tree dependence should also be considered with similar care when attempting to reconstruct species trees.

Our findings are particularly relevant as whole-genome sequencing produces more gene trees for each genome. It is tempting to think that increasing the number of gene trees will counter the inaccuracy resulting from gene tree dependence. However, the number of genes in a genome is not unlimited, and furthermore, including more gene trees will lower the genomic distance between the trees, and thus increase the dependence between them, making the effect of gene tree dependence more pronounced. Thus, as more and more data is collected, it becomes more important to consider the effect of gene tree dependence. For the relatively small species tree we study here, we can achieve a satisfactory reconstruction accuracy with relatively few gene trees even when dependence is taken into account, but this will not necessarily always be the case. For larger trees that are more difficult to reconstruct exactly, the number of gene trees needed to achieve a satisfactory accuracy may well exceed the actual number of genes available.

Although we have restricted our simulations to ASTRAL as it is currently the most popular summary method for species tree inference, it is likely that many other summary methods (in particular, methods that attempt to solve the maximum quartet support problem as ASTRAL does) will be similarly affected by gene tree dependence. Virtually all simulations evaluating these methods are performed with independent gene trees, and therefore significantly over-estimate the accuracy of these methods.

We still assume that recombination only happens between loci, but it is also possible to have recombination within loci (intra-locus recombination). Although it has been shown that intra-locus recombination has little effect on species tree inference (Lanier and Knowles, 2012; Zhu et al., 2022), it has also been shown (Wang and Liu, 2016) that proper breakpoint identification can improve species tree inference. The traditional consideration of a gene as a non-recombinant locus introduces potential complications in that adjacent trees are very closely related, increasing the effective amount of recombination between loci. An alternative treatment of this may involve gene networks rather than trees, such as in Chan and Robin (2019), which significantly complicates the problem.

To estimate a realistic recombination rate for the biological dataset, we used the average consecutive pairwise normalised RF distance between gene trees. Other methods to calibrate the recombination rate could be explored; for example, a direct way is to reconstruct the ARG with recombination events and estimate the recombination rate from there. Indeed, one would expect that the recombination rate would vary across datasets, depending in particular on the density of the genes within the genome, and would need to be re-estimated for each specific case. Given that genes can also vary in position between species, it is not clear that there even is one 'true' recombination rate; the recombination rate that we calculate here should be considered as an overall average, incorporating all the above effects.

When estimating the the effective sample size for ASTRAL, there is an accumulated error including the estimation of recombination rate, the accuracy of different sample sizes, and the

matching of sample sizes to a given accuracy. Future work includes proposing a simpler and more reliable method to estimate the effective sample size.

Although we have shown gene tree dependence to have a detrimental effect on the accuracy of ASTRAL, it is unclear how to counter this phenomenon. We show that thinning (subsampling) gene trees does not help, and by extension neither does selection of loci; the best accuracy is obtained by using the most trees even if they are more dependent. At present, ASTRAL is still the most effective summary method available; our results merely present a note of caution about how accurate it truly is on real data.

Of course, this suggests that a new species tree inference method to take gene tree dependence into account should be designed. One possible way is to assign different weights to quartets according to their patterns of occurrence in the genome. By incorporating the information of genomic location, this new method could potentially be more accurate in the case of gene tree dependence. This is the subject of future work.

Fundings

Celine Scornavacca was supported by French Agence Nationale de la Recherche through CoCoAlSeq project (ANR-19-CE45-0012).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data and code availability

Data and code are available on https://github.com/WAHE1/ASTI-GTD/tree/main.

References

Ballesteros JA, Sharma PP (2019). *A critical appraisal of the placement of* Xiphosura *(*Chelicerata*) with account of known sources of phylogenetic error*. *Systematic Biology* **68**, 896–917.

Barker D, Pagel M (2005). *Predicting Functional Gene Links from Phylogenetic-Statistical Analyses of Whole Genomes*. *PLoS Computational Biology* **1**, 1–8.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, Gladstein AL, Gorjanc G, Guo B, Jeffery B, Kretzschumar WW, Lohse K, Matschiner M, Nelson D, Pope NS, Quinto-Cortés CD, et al. (2022). *Efficient ancestry and mutation simulation with msprime 1.0*. *Genetics* **220**, iyab229.

Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, Rheinbay E, Grabherr M, Forche A, Reedy JL, et al. (2009). *Evolution of pathogenicity and sexual reproduction in eight* Candida *genomes*. *Nature* **459**, 657–662.

Chan Y, Li Q, Scornavacca C (2022). *The large-sample asymptotic behaviour of quartet-based summary methods for species tree inference*. *Journal of Mathematical Biology* **85**, 1–22.

Chan Y, Robin C (2019). *Reconciliation of a gene network and species tree*. *Journal of Theoretical Biology* **472**, 54–66.

Conry M (2020). *Determining the impact of recombination on phylogenetic inference*. The Florida State University.

Davidson R, Vachaspati P, Mirarab S, Warnow T (2015). *Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer*. *BMC Genomics* **16**, 1–12.

DeGiorgio M, Degnan JH (2014). *Robustness to divergence time underestimation when inferring species trees from estimated gene trees*. *Systematic Biology* **63**, 66–82.

Felsenstein J (2004). *Inferring phylogenies*. In: *Inferring phylogenies*, pp. 664–664.

Giarla TC, Esselstyn JA (2015). *The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews*. *Systematic Biology* **64**, 727–740.

Griffiths RC, Marjoram P (1997). *An ancestral recombination graph*. *Institute for Mathematics and its Applications* **87**, 257.

Hasegawa M, Kishino H, Yano Ta (1985). *Dating of the human-ape splitting by a molecular clock of mitochondrial DNA*. *Journal of Molecular Evolution* **22**, 160–174.

Huang H, Knowles LL (2016). *Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences*. *Systematic Biology* **65**, 357–365.

Hudson RR (1983). *Properties of a neutral allele model with intragenic recombination*. *Theoretical Population Biology* **23**, 183–201.

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

Kelleher J, Etheridge AM, McVean G (2016). *Efficient coalescent simulation and genealogical analysis for large sample sizes*. *PLoS Computational Biology* **12**, e1004842.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G (2019). *Inferring whole-genome histories in large population datasets*. *Nature Genetics* **51**, 1330–1338.

Kingman JF (1982). *On the genealogy of large populations*. *Journal of Applied Probability* **19**, 27–43.

Lanier HC, Knowles LL (2012). *Is recombination a problem for species-tree analyses?* *Systematic Biology* **61**, 691–701.

Lanier HC, Knowles LL (2015). *Applying species-tree analyses to deep phylogenetic histories: challenges and potential suggested from a survey of empirical phylogenetic studies*. *Molecular Phylogenetics and Evolution* **83**, 191–199.

Legried B, Molloy EK, Warnow T, Roch S (2021). *Polynomial-time statistical estimation of species trees under gene duplication and loss*. *Journal of Computational Biology* **28**, 452–468.

Li Q, Scornavacca C, Galtier N, Chan Y (2021). *The multilocus multispecies coalescent: a flexible new model of gene family evolution*. *Systematic Biology* **70**, 822–837.

Liu L, Yu L (2011). *Estimating species trees from unrooted gene trees*. *Systematic Biology* **60**, 661–667.

Liu L, Yu L, Edwards SV (2010). *A maximum pseudo-likelihood approach for estimating species trees under the coalescent model*. *BMC Evolutionary Biology* **10**, 1–18.

Liu L, Yu L, Pearl DK, Edwards SV (2009). *Estimating species phylogenies using coalescence times among sequences*. *Systematic Biology* **58**, 468–477.

Maddison WP (1997). *Gene trees and species trees*. *Systematic Biology* **46**, 523–536.

Mahmoudi A, Koskela J, Kelleher J, Chan Y, Balding D (2022). *Bayesian inference of ancestral recombination graphs*. *PLOS Computational Biology* **18**, e1009960.

Markin A, Eulenstein O (2021). *Quartet-based inference is statistically consistent under the unified duplication-loss-coalescence model*. *Bioinformatics* **37**, 4064–4074.

McVean GA, Cardin NJ (2005). *Approximating the coalescent with recombination*. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360**, 1387–1393.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, Lanfear R (2020). *IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era*. *Molecular Biology and Evolution* **37**, 1530–1534.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T (2014). *ASTRAL: genome-scale coalescent-based species tree estimation*. *Bioinformatics* **30**, i541–i548.

Mirarab S, Warnow T (2015). *ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes*. *Bioinformatics* **31**, i44–i52.

Molloy EK, Warnow T (2018). *To include or not to include: the impact of gene filtering on species tree estimation methods*. *Systematic Biology* **67**, 285–303.

Mossel E, Roch S (2008). *Incomplete lineage sorting: consistent phylogeny estimation from multiple loci*. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**, 166–171.

Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ (2015). *IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-Likelihood phylogenies*. *Molecular Biology and Evolution* **32**, 268–274.

Pamilo P, Nei M (1988). *Relationships between gene trees and species trees*. *Molecular Biology and Evolution* **5**, 568–583.

Patel S, Kimball RT, Braun EL (2013). *Error in phylogenetic estimation for bushes in the tree of life*. *Journal of Phylogenetics and Evolutionary Biology* **1**, 1–10.

Price MN, Dehal PS, Arkin AP (2009). *FastTree: computing large minimum evolution trees with profiles instead of a distance matrix*. *Molecular Biology and Evolution* **26**, 1641–1650.

Price MN, Dehal PS, Arkin AP (2010). *FastTree 2–approximately maximum-likelihood trees for large alignments*. *PLOS ONE* **5**, e9490.

Rannala B, Yang Z (2003). *Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci*. *Genetics* **164**, 1645–1656.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014). *Genome-wide inference of ancestral recombination graphs*. *PLOS Genetics* **10**, e1004342.

Rasmussen MD, Kellis M (2012). *Unified modeling of gene duplication, loss, and coalescence using a locus tree*. *Genome Research* **22**, 755–765.

Robinson DF, Foulds LR (1981). *Comparison of phylogenetic trees*. *Mathematical Biosciences* **53**, 131–147.

Sayyari E, Mirarab S (2016). *Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction*. *BMC Genomics* **17**, 101–113.

Slatkin M, Pollack JL (2006). *The concordance of gene trees and species trees at two linked loci*. *Genetics* **172**, 1979–1984.

Song S, Liu L, Edwards SV, Wu S (2012). *Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model*. *Proceedings of the National Academy of Sciences* **109**, 14942–14947.

Speidel L, Forest M, Shi S, Myers SR (2019). *A method for genome-wide genealogy estimation for thousands of samples*. *Nature Genetics* **51**, 1321–1329.

Spielman SJ, Wilke CO (2015). *Pyvolve: a flexible Python module for simulating sequences along phylogenies*. *PloS one* **10**, e0139047.

Tavaré S (1986). *Some probabilistic and statistical problems on the analysis of DNA sequence*. *Lecture of Mathematics for Life Science* **17**, 57.

Vachaspati P, Warnow T (2015). *ASTRID: accurate species trees from internode distances*. *BMC Genomics* **16**, 1–13.

Wang Z, Liu KJ (2016). *A performance study of the impact of recombination on species tree analysis*. *BMC Genomics* **17**, 165–174.

Wu YC, Rasmussen MD, Bansal MS, Kellis M (2014). *Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees*. *Genome Research* **24**, 475–486.

Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L (2022). *Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs*. *Systematic Biology* **71**, 367–381.

Yin J, Zhang C, Mirarab S (2019). *ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization*. *Bioinformatics* **35**, 3961–3969.

Zhang C, Rabiee M, Sayyari E, Mirarab S (2018). *ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees*. *BMC Bioinformatics* **19**, 15–30.

Zhu T, Flouri T, Yang Z (2022). *A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model*. *Molecular Ecology* **31**, 2814–2829.

Supporting information

Re-estimating branch lengths for the 37-taxon mammalian tree

The terminal branch lengths of the species tree given in Mirarab et al. (2014) are all 9.0 coalescent units; this is an artifact of the MP-EST summary method used to estimate it. This is immaterial for the topologies of the gene trees simulated under the MSC (or any related model) from this species tree; however, it has an effect on the branch lengths of these trees, and therefore on sequences generated from these trees.

To amend this, we re-estimate the branch lengths for the 37-taxon mammalian tree from the original data of Song et al. (2012). We concatenate the sequences of the 447 genes from that paper, and use them as input to IQ-TREE (Minh et al., 2020; Nguyen et al., 2015) while constraining the topology of the tree to that used in Mirarab et al. (2014). The resulting tree is then re-rooted on the leaf branch 'Gal' by splitting the branch into two halves such that the length of the branch to 'Gal' is the average of the lengths of the paths from the root to the other leaf nodes.

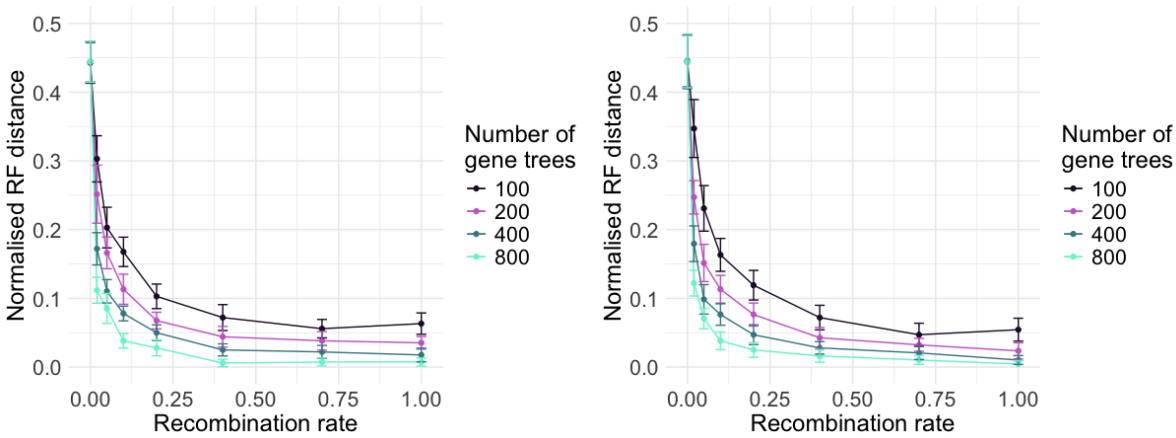
The branch lengths of the resulting tree are given in units of (average number of) substitutions per site. To translate from this unit to coalescent units (in the absence of estimated mutation rates and population sizes), we scale the tree so that the average internal branch length is equal to that of the tree in Mirarab et al. (2014).

To verify that our gene trees have more realistic branch lengths than the trees from Mirarab et al. (2014), we simulated 4000 gene trees from the MSC model on our species tree, and scaled the branch lengths by a simple multiplier to produce an average branch length equal to that estimated from the dataset (Song et al., 2012). We then calculated the weighted Robinson-Foulds distance (Felsenstein, 2004) between each simulated gene tree and each biological tree from Song et al. (2012), and did the same for the gene trees from Mirarab et al. (2014), which were simulated from their species tree and re-scaled in a different way. The average weighted RF distance for our trees is 1.896, while the average distance for trees using the Mirarab et al. (2014) procedure is 2.141, which is significantly higher (p-value $< 2.2 \times 10^{-16}$ from a two-sample t-test). This suggests that our gene trees are more realistic.

Supplementary figures

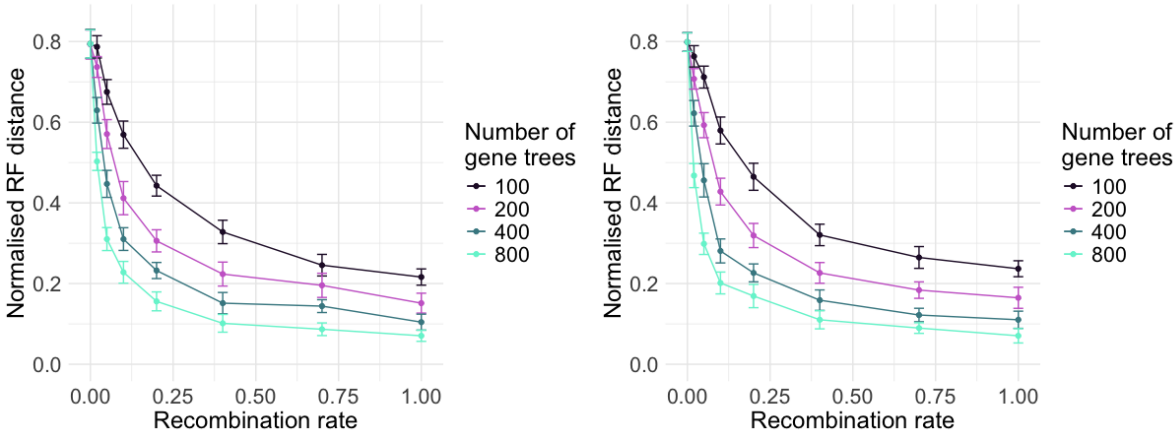
Species name	Abbreviation	Species name	Abbreviation
Vicugna pacos	VIC	*Pongo pygmaeus*	PON
Bos taurus	BOS	*Tarsius syrichta*	TAR
Sus scrofa	SUS	*Dipodomys ordii*	DIP
Equus caballus	EQU	*Mus musculus*	MUS
Canis familiaris	CAN	*Rattus*	RAT
Felis catus	FEL	*Spermophilus tridecemlineatus*	SPE
Tursiops truncatus	TUR	*Cavia porcellus*	CAV
Erinaceus europaeus	ERI	*Ochotona princeps*	OCH
Pteropus vampyrus	PTE	*Oryctolagus cuniculus*	ORY
Myotis lucifugus	MYO	*Loxodonta africana*	LOX
Sorex araneus	SOR	*Procavia capensis*	PRO
Tupaia belangeri	TUP	*Echinops telfairi*	ECH
Otolemur garnettii	OTO	*Dasypus novemcinctus*	DAS
Pan troglodytes	PAN	*Choloepus hoffmanni*	CHO
Gorilla gorilla	GOR	*Monodelphis domestica*	MON
Homo sapiens	HOM	*Macropus eugenii*	MAC
Macaca mulatta	MAC	*Ornithorhynchus anatinus*	ORN
Callithrix jacchus	CAL	*Gallus gallus*	GAL
Microcebus murinus	MIC		

Figure 15 – Abbreviations of species names in the 37-taxon mammalian species tree in Fig 3.



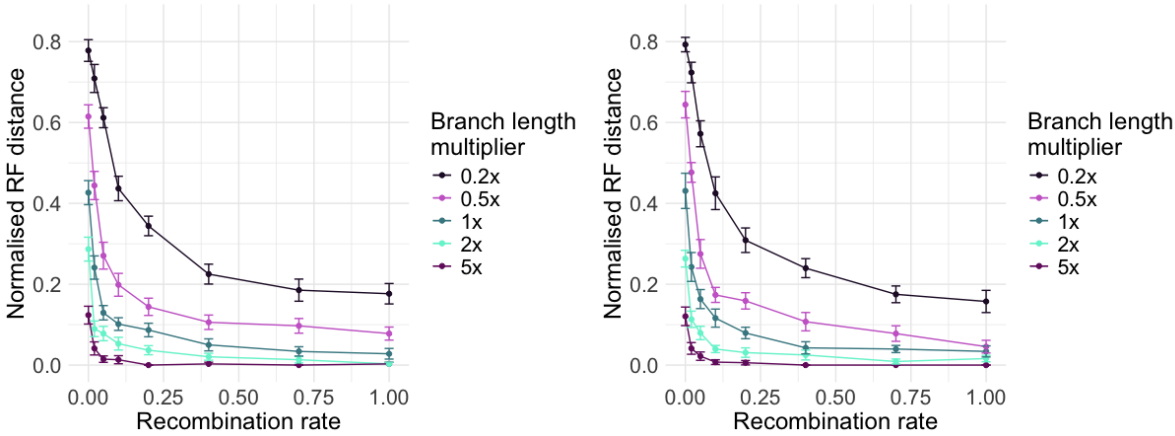
(a) 500bp sequence length **(b)** 1000bp sequence length

Figure 16 – The accuracy of ASTRAL vs recombination rate for estimated gene trees, averaged over 20 replicates. Error bars represent standard error.



(a) 500bp sequence length **(b)** 1000bp sequence length

Figure 17 – The accuracy of ASTRAL vs recombination rate with increased ILS ($0.2\times$) for estimated gene trees, averaged over 20 replicates. Error bars represent standard error.



(a) 500bp sequence length **(b)** 1000bp sequence length

Figure 18 – The accuracy of ASTRAL vs recombination rate with different amount of ILS for estimated gene trees, averaged over 20 replicates. Error bars represent standard error.

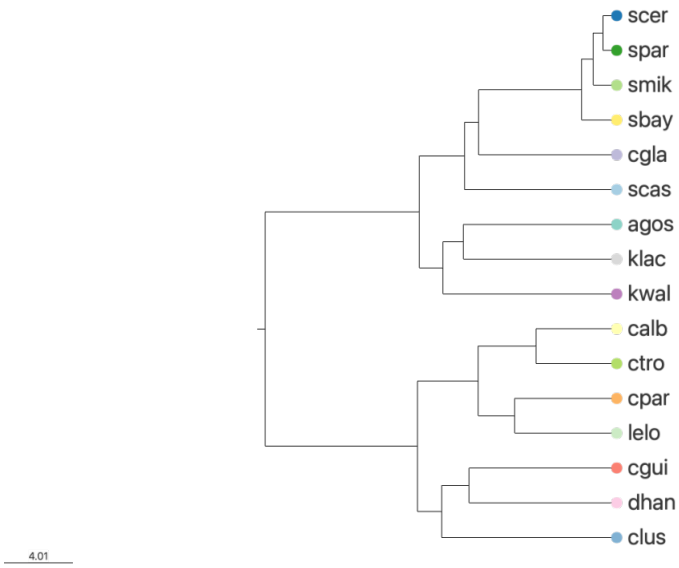


Figure 19 – The 16-taxon fungi species tree with branch lengths in coalescent units. Species names are given in Table 20 in Supporting information.

Species name	Abbreviation
Candida albicans	calb
Candida tropicalis	ctro
Candida parapsilosis	cpar
Lodderomyces elongisporus	lelo
Candida guilliermondii	cgui
Debaryomyces hansenii	dhan
Candida lusitaniae	clus
Saccharomyces cerevisiae	scer
Saccharomyces paradoxus	spar
Saccharomyces mikatae	smik
Saccharomyces bayanus	sbay
Candida glabrata	cgla
Saccharomyces castellii	scas
Kluyveromyces latics	klac
Ashbya gossypii	agos
Kluyveromyces waltii	kwal

Figure 20 – Abbreviations of species names in the 16-taxon fungi species tree in Figure 19 in Supporting information.

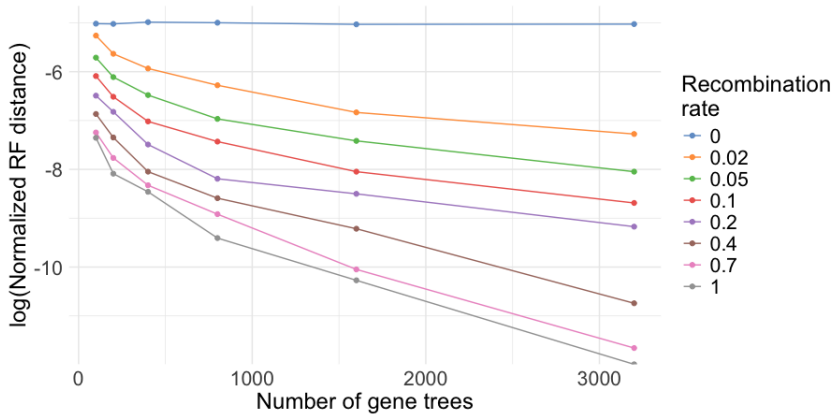
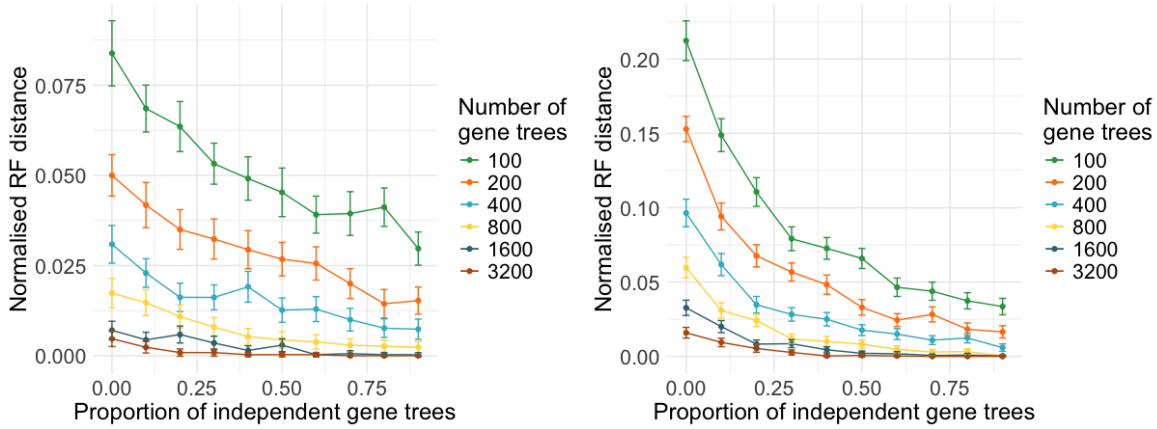
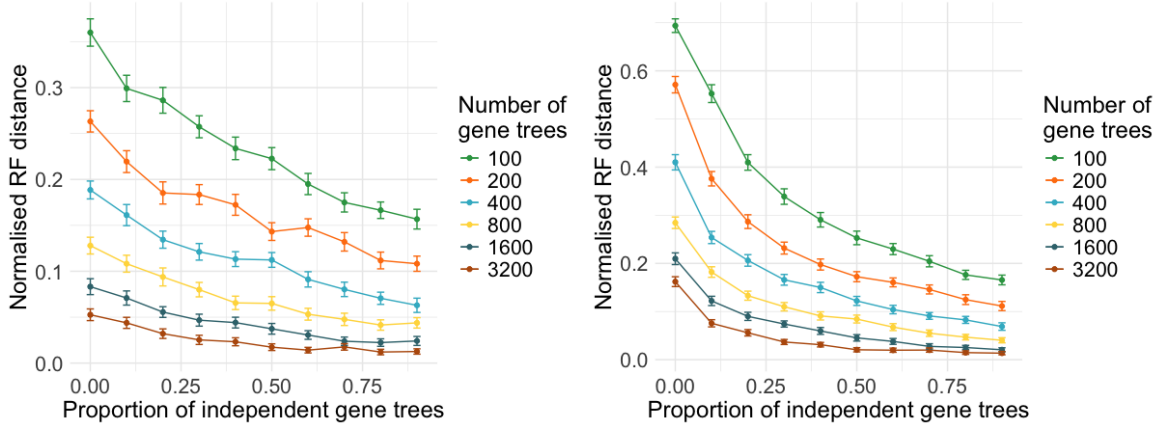


Figure 21 – The accuracy of ASTRAL (log normalised RF distance) vs the number of gene trees, averaged over 1000 replicates.



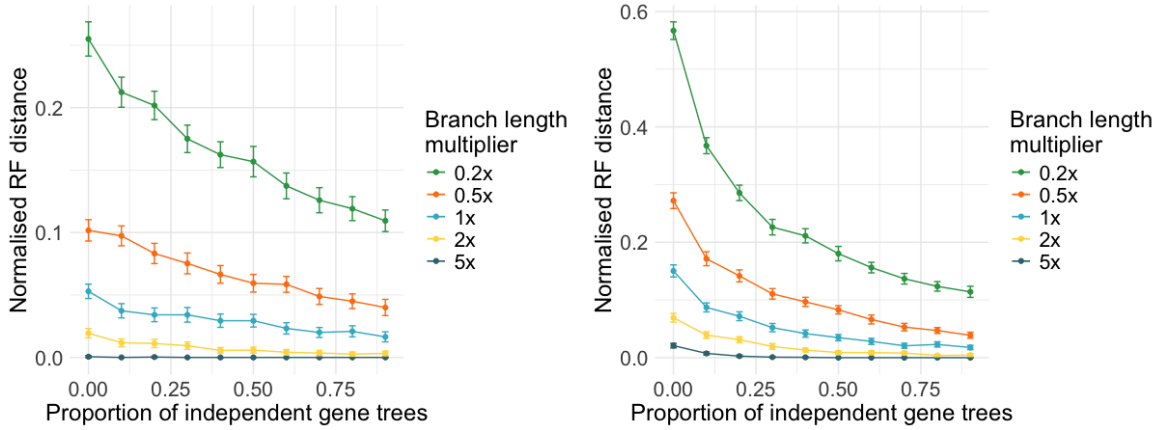
(a) Default for $R = 0.317$

(b) Default for $R = 0.055$



(c) Increased ILS (0.2×) for $R = 0.317$

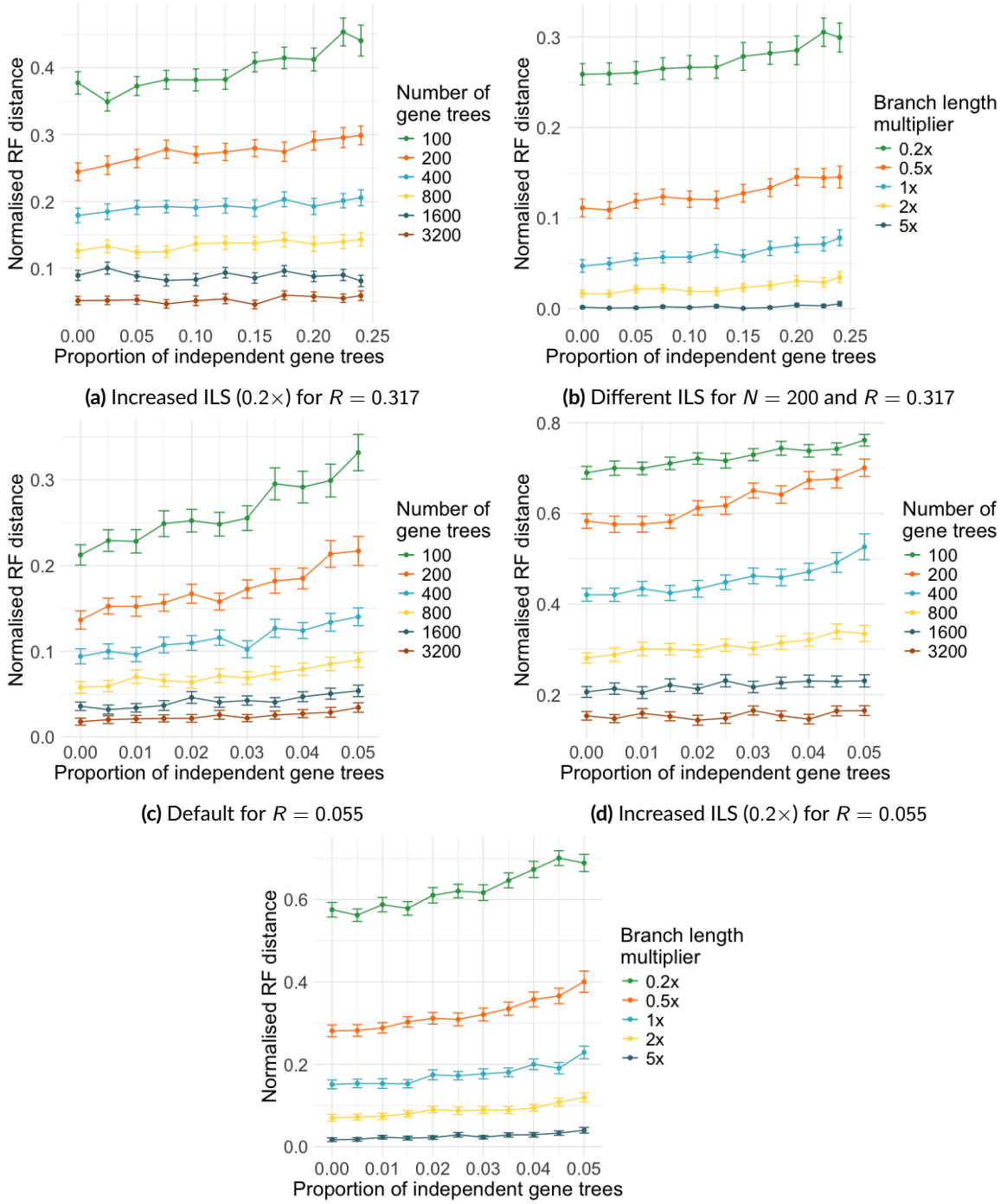
(d) Increased ILS (0.2×) for $R = 0.055$



(e) Different ILS for $N = 200$ and $R = 0.317$

(f) Different ILS for $N = 200$ and $R = 0.055$

Figure 22 – The accuracy of ASTRAL vs the proportion of independent gene trees, averaged over 100 replicates. Error bars represent standard error.



(a) Increased ILS (0.2×) for $R = 0.317$

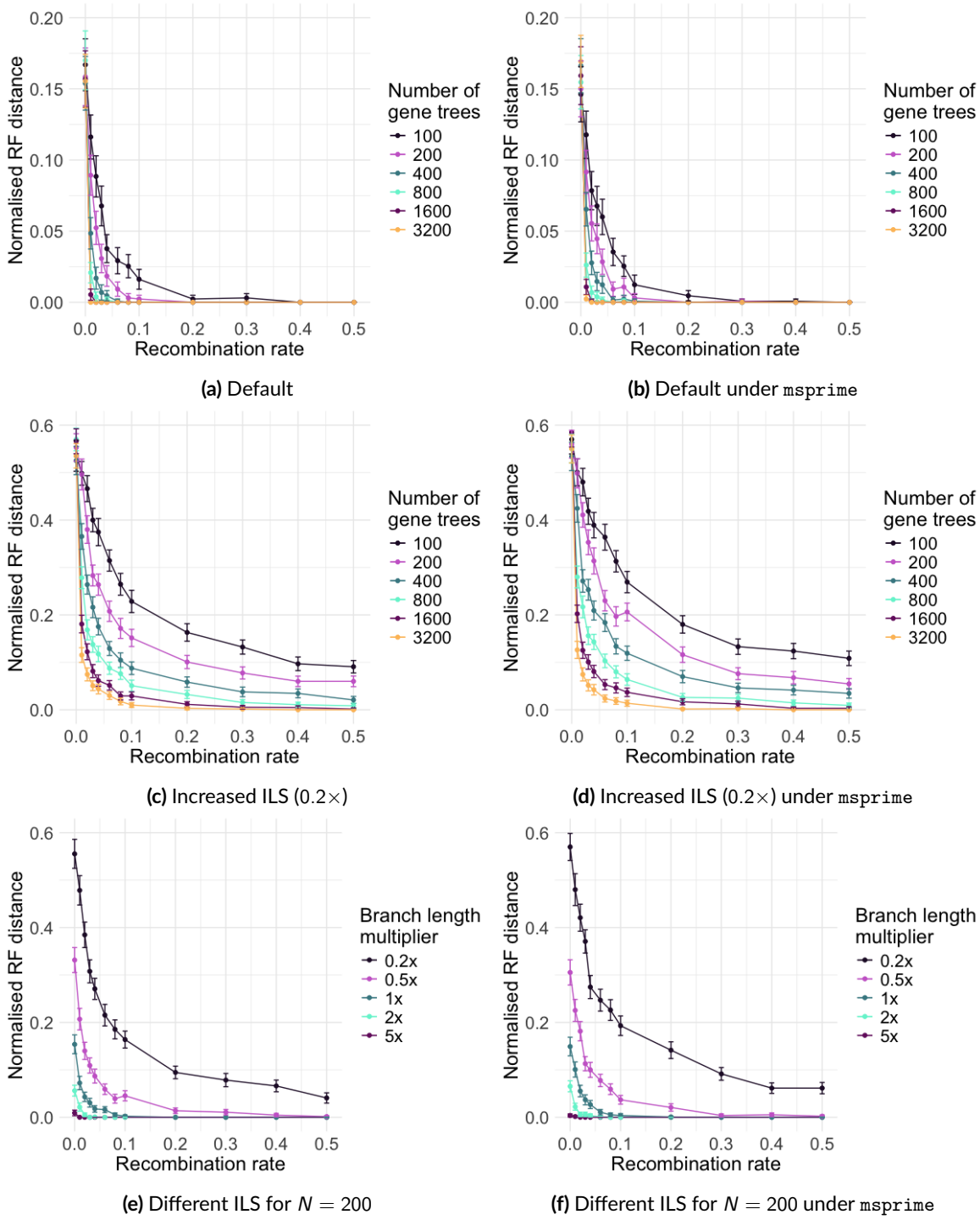
(b) Different ILS for $N = 200$ and $R = 0.317$

(c) Default for $R = 0.055$

(d) Increased ILS (0.2×) for $R = 0.055$

(e) Different ILS for $N = 200$ and $R = 0.055$

Figure 23 – The accuracy of ASTRAL vs the proportion of independent gene trees, averaged over 100 replicates, with fixed overall recombination rate. Error bars represent standard error. The default settings are shown in Figure 11.



(a) Default

(b) Default under `msprime`

(c) Increased ILS (0.2×)

(d) Increased ILS (0.2×) under `msprime`

(e) Different ILS for $N = 200$

(f) Different ILS for $N = 200$ under `msprime`

Figure 24 – The accuracy of ASTRAL vs recombination rate for the 16-taxon fungi species tree, averaged over 100 replicates. Error bars represent standard error.

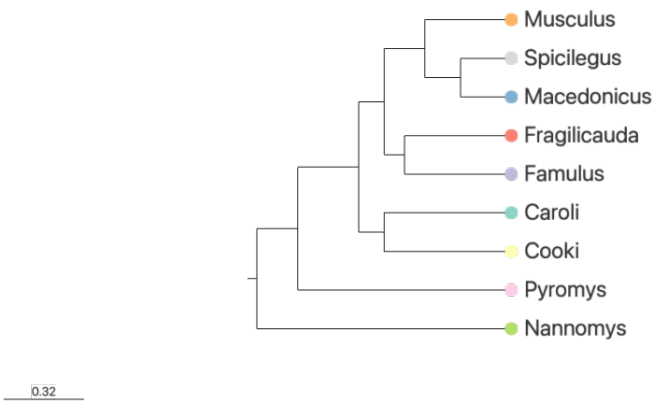


Figure 25 – The 9-taxon mouse species tree from Wang and Liu (2016).