The accuracy of species tree inference under gene tree dependence

Wanting He[1,2], Celine Scornavacca[3], Yao-ban Chan[1,2*],

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

* yaoban@unimelb.edu.au

Abstract

When inferring the evolutionary history of species and the genes they contain, the phylogenetic trees of the genes can be different to that 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 the 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 related by function) 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 larger 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 Robinson-Foulds distance increases by a factor of 4.7) relative to the accuracy previously estimated with independent gene trees, and the effective sample size for this dataset is about one-third 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 extensive.

Author summary

The study of the evolutionary history of species is important for understanding and reconstructing the history of life on Earth. These evolutionary histories are represented in the form of species trees, which can be reconstructed from the evolutionary histories of the genes contained in the species using so-called species tree inference methods. This is complicated by the fact that the histories of the genes (gene trees) can be related to

each other, depending on their placement in the genome or their functions. Gene tree dependence is not taken into account in almost all studies of the accuracy of species tree inference. In this paper, we develop a statistical model to include gene tree dependence, and show that it can significantly affect the accuracy of species tree inference. This effect is at least as large as the impact of incorrect reconstruction of the gene trees themselves, a well-known issue in species tree inference.

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, the gene copies within these species also diverge. Therefore, gene trees can be considered as contained 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) [1]. In particular, ILS, where multiple gene lineages fail to coalesce over a series of multiple speciations, is a major cause of discordance between the gene and species trees (see Fig 1 for an example). The standard statistical model for gene trees under ILS is the multispecies coalescent (MSC) model [2, 3], in which each species branch represents a separate population for which Kingman's basic coalescent [4] is run in a bottom-up fashion where lineages at the top of each branch are used as input to their parent branches.

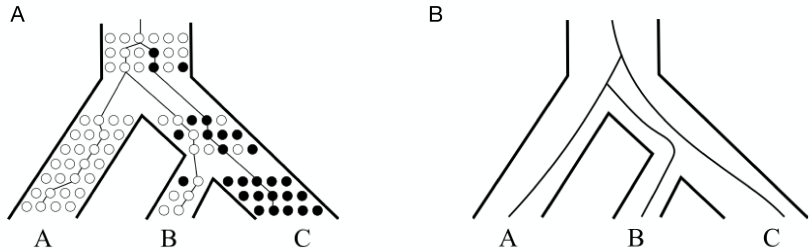

Fig 1. An example of incomplete lineage sorting. Here, circles are individuals of a population and each row is a generation. Originally, the population contains only empty circles, but a filled circle was generated by a mutation. After the first speciation, the empty circles fix in species A, but the gene remains in a polymorphic state in the ancestor of B and C, and subsequently the empty circles fix in B and the filled circles in C. As a result, 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. **A**: Population-level view. **B**: Gene tree view.

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 supergene, and then a species tree is obtained by applying traditional phylogenetic inference methods, such as parsimony and maximum likelihood. This paradigm 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 [5], NJst [6], ASTRID [7], DISTIQUE [8], ASTRAL [9–11] and STAR [12], and some use both gene tree topologies and branch lengths, such as GLASS [13] and STEAC [12]. 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 [14, 15] and scalability [10, 16]. It has been shown to give a statistically consistent estimator of the species tree if input gene trees are sampled under the MSC model [9], bounded HGT models [17], the general DLCoal model [18], or the GDL model [19, 20]. Furthermore, extensive simulations have been performed studying its accuracy under practical conditions, showing that ASTRAL is highly accurate under the MSC model [7–10], the SimPhy model [17], and with gene filtering when ILS is low to moderate [21]. 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 [21–25].

Although the accuracy of ASTRAL has been extensively studied by simulations, almost all simulations were performed with gene trees that were independently simulated 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 [26, 27] 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 relationships, may also induce dependence between gene trees. 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 [28], 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 [29] 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 [30]. This produces an evolutionary history which contains both coalescent and recombination events, known as the ancestral recombination graph (ARG) [31]. The problem of inferring the ARG from genome sequences has been extensively studied, with great progress in recent years [32–35]. Earlier methods, such as ARGWeaver [32], achieved efficiency by using an approximation to the coalescent with recombination known as the sequentially Markov coalescent [36], 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. We extend the two-locus model of [37] to multiple species and loci to derive a probabilistic model for generating dependent gene trees. We then use the generated trees as input to ASTRAL to estimate the species tree. When re-analysing a 37-taxon mammalian dataset [9], we find that under realistic recombination rates, the accuracy of ASTRAL may have been significantly over-estimated, with an effective sample size roughly one-third that of the actual sample size. This indicates that gene tree dependence may be a significant factor when estimating the accuracy, potentially even more so than gene tree estimation error.

Methods

We first describe a model 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 two-locus model of Slatkin and Pollack [37], who devised a model of two linked loci in three species based on the coalescent with recombination. They used it to calculate the probability that these gene trees are concordant with the species tree and with each other. This was also used in [38] to model so-called 'linked duplications'. Here, we first consider two loci of haploid individuals in a single Wright-Fisher (panmictic) population, and we are interested in the genealogy of samples in the second locus, conditional on their genealogy in the first locus. Fig 2 shows an example of our model (described below).

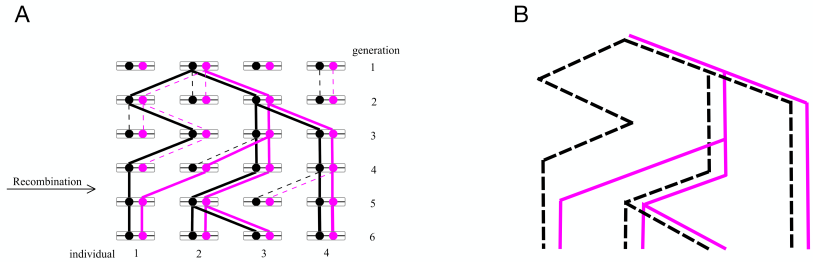

Fig 2. An example of the two-locus model. The black and pink lines represent the genealogies in the first and second loci respectively. **A**: Population view. A recombination event happens in the leftmost individual of generation 5, which leads to a divergence of lineages between the two loci, and thus 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.

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 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. 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 [36], sampled at constant intervals along the genome.

This 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 vary the number of blocks while fixing the average value of $\frac{R}{1+R}$ between trees. An intuitive explanation for this expression is that lineages coalesce with 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 the 37-taxon mammalian species tree shown in Fig 3, which has previously been used to study the accuracy of ASTRAL for independent gene trees [9]. The species tree was previously estimated with MP-EST [5] on the biological dataset from Song *et al.* [39], containing 447 genes with average length 3099 bp. In order to reassess the accuracy of ASTRAL with dependent gene trees, we perform simulations with the same parameter settings as in [9].

Given a species tree, we generate a set of N dependent gene trees using our model above. We then use these trees as input to ASTRAL to infer a species tree, and calculate the Robinson-Foulds (RF) distance [40] between the true and inferred species tree as a measure of the accuracy of ASTRAL.

Additionally, we rescale the simulated dependent gene trees to produce average branch lengths equal to those estimated from the dataset [39], then evolve gene sequences along the trees with `Pyvolve` [41], under the GTR model of site evolution [42]

with parameters estimated from the biological dataset. As in [9], we choose half of the sequences to be 500 bp and half to be 1000 bp. We then use IQ-TREE [43, 44] 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 2, 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.

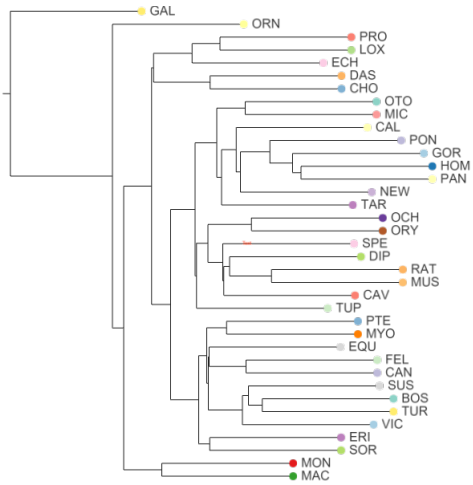

Fig 3. The 37-taxon mammalian species tree (see species names in Table 14 in Supporting information).

It has been shown that ASTRAL performs worse with an increased amount of ILS [9], 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×), 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× to 5× as in [9].

The parameter settings for the simulations are summarised in Table 1. We perform 1000 replicates for each parameter setting for true gene trees, and 100 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 for the alternative way of generating dependent gene trees, 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.

Table 1. Simulation parameters.

Simulation	R	N	**ILS**[1]
Default simulations	[0, 2]	[100, 3200]	1× branch lengths
Increased amount of ILS	[0, 2]	[100, 1600]	0.2× branch lengths
Different amounts of ILS	[0, 2]	200	[0.2, 5]× branch lengths

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

To verify that the our results are not biased by the specific species tree used, we repeat the above simulations on a tree of 16 fungi [18, 19, 45–47], shown in Fig 15 in

Supporting information. This fungal tree has been studied extensively in simulations aiming to show that the accuracy of ASTRAL is not affected by data with paralogs [20].

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.

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 inferred gene trees in Fig 4.

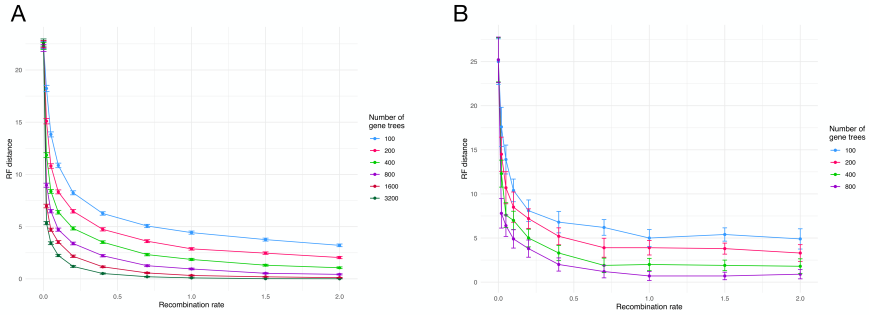

Fig 4. The accuracy of ASTRAL vs recombination rate, averaged over a number of replicates. Error bars represent standard error. **A**: True dependent gene trees (1000 replicates). **B**: Estimated dependent gene trees (100 replicates).

In the Fig 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 RF distance decreases — ASTRAL becomes more accurate with less dependence among gene trees. As $R \to \infty$, the accuracy converges to the results of [9] 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 [48] that with independent gene trees, the RF distance decreases exponentially with respect to N. In Fig 17 in Supporting information, we show a log-plot of 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.

In Fig 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 Fig 5, we compare the relative effects of these factors with recombination rate fixed to a realistic value (estimated in Estimating the recombination rate) of 1.26. We can see that for this recombination rate, the loss in accuracy from gene tree dependence is considerably larger than the corresponding loss from gene tree estimation error.

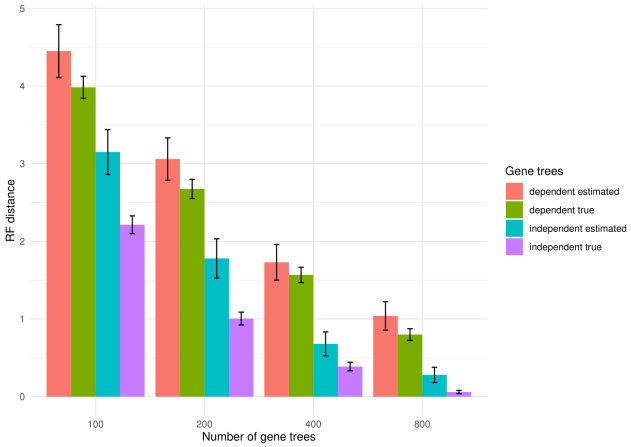

Fig 5. The accuracy of ASTRAL with and without gene tree estimation error and dependence. The accuracy is shown against the number of gene trees for $\boldsymbol{R = 1.26}$. Error bars represent standard error.

Increased amount of ILS

We repeat our simulations above with an increased amount of ILS (0.2× branch lengths). We observe similar patterns to Default simulations, as shown in Fig 6. 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.

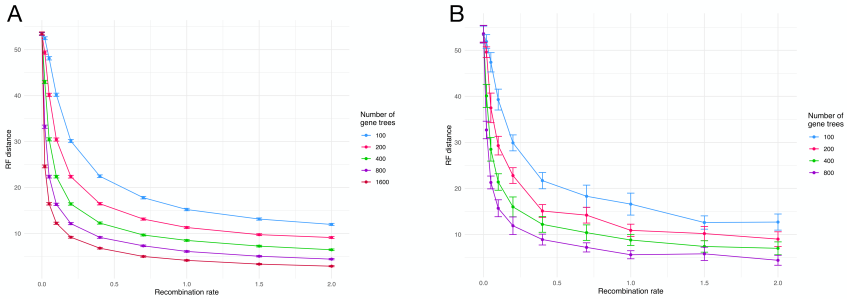

Fig 6. The accuracy of ASTRAL vs recombination rate with increased ILS (0.2×). Error bars represent standard error. **A**: True dependent gene trees (1000 replicates). **B**: Estimated dependent gene trees (100 replicates).

In Fig 7, we consider the relative effects of gene tree dependence and estimation error under increased amounts of ILS and a realistic recombination rate $R = 1.26$. 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.

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 Fig 8, we fix the number of gene trees to 200 and vary the branch length multiplier. We observe similar patterns of RF distance with respect to recombination rate as above. Overall, ASTRAL performs better with less ILS, as observed for independent gene trees in [9].

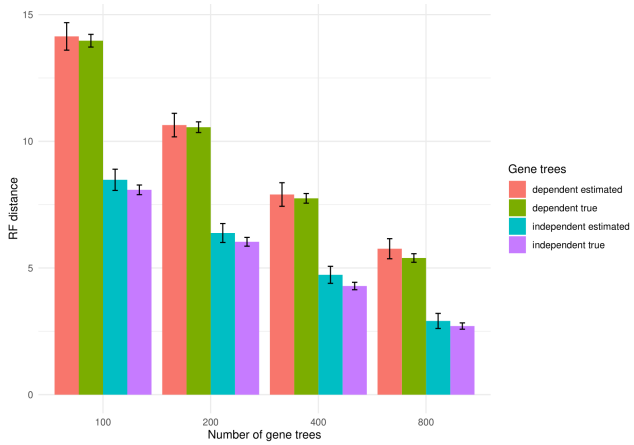

Fig 7. The accuracy of ASTRAL with and without gene tree estimation error and dependence, with increased ILS (0.2×). The accuracy is shown against the number of gene trees for $\boldsymbol{R = 1.26}$. Error bars represent standard error.

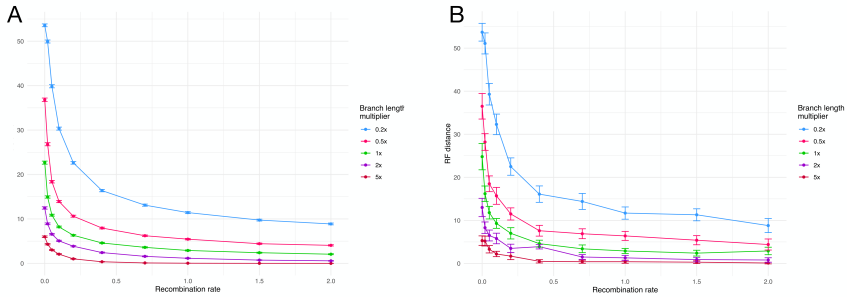

Fig 8. The accuracy of ASTRAL vs recombination rate with different amounts of ILS for $\boldsymbol{N = 200}$. Error bars represent standard error. **A**: True dependent gene trees (1000 replicates). **B**: Estimated dependent gene trees (100 replicates).

In Fig 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. Only when branch lengths are scaled by a factor of 5 does gene tree estimation error have a larger overall effect.

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 Generating dependent gene trees. 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 Fig 18 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 Fig 10 (see also Fig 19 in Supporting information), we fix the overall recombination rate to be equivalent to the uniform dependence case with $R = 1.26$. We

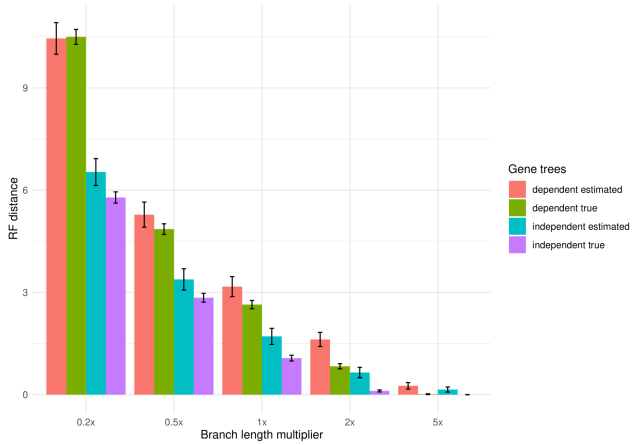

Fig 9. The accuracy of ASTRAL with and without gene tree estimation error and dependence, with different amounts of ILS. The accuracy is shown against the number of gene trees for $\boldsymbol{R = 1.26}$. Error bars represent standard error.

observe that the dependence structure does not significantly affect the accuracy of ASTRAL: more dependence within smaller blocks makes no difference to 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.

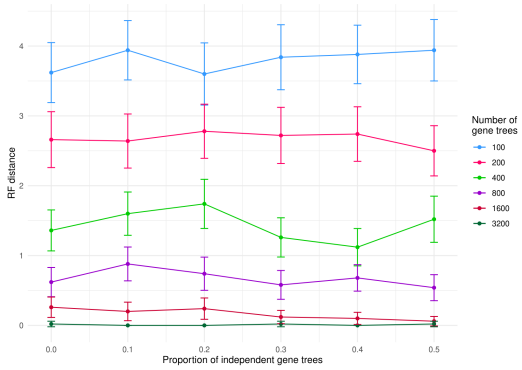

Fig 10. The accuracy of ASTRAL vs the proportion of independent gene trees. The accuracy is averaged over 100 replicates, with fixed overall recombination rate. Error bars represent standard error.

16-taxon fungal species tree

In Fig 20 in Supporting information, we observe similar patterns to those in the 37-taxon mammalian species tree, but with overall higher accuracy. This indicates that our results are not specific to the mammalian species tree.

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.* [39] on the 37-taxon mammalian species tree studied above. This dataset contains 424 effective genes (after 21 genes with mislabelled

sequences and 23 outliers were removed from the original 477-gene dataset) with average length 3157 bp. We re-estimated the gene trees with IQ-TREE [43, 44].

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 RF distance between the real trees to those from simulations. The average consecutive pairwise RF distance of the real gene trees is 24.913; this is significantly different from the value for simulated independent gene trees on the same species tree (31.504; 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 RF distance. For a range of recombination rates from 0 to 10, we simulate the same number of gene trees, and calculate the average consecutive pairwise RF distance. The results are shown in Fig 11. To estimate the recombination rate, we repeat a binary search 100 times on the range $[0, 2]$ to reduce sampling variation, and estimate $R = 1.26$ with a standard deviation of 0.002.

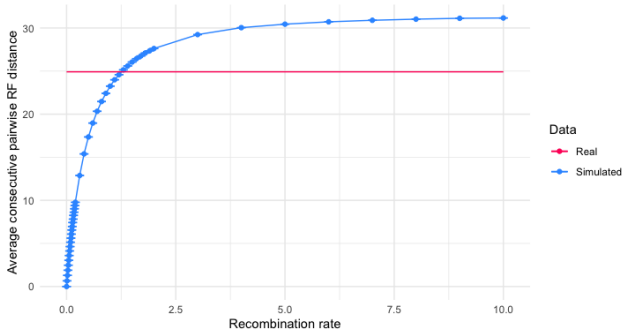

Fig 11. The average consecutive pairwise RF distance between gene trees. Error bars represent standard error.

From our simulations, the accuracy of ASTRAL with 424 independent genes is 0.308 average RF distance with standard error 0.024. However, under this recombination rate, the accuracy is approximately 1.452 average RF distance with standard error 0.047. Thus, by ignoring dependence between gene trees, we are overestimating the accuracy of ASTRAL for this dataset by a factor of 4.7.

Effective sample size

The multiplicative decrease in 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 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 RF distance for

ASTRAL for the estimated recombination rate $R = 1.26$. We then use a binary search to determine the number of independent gene trees required to produce the same RF distance. For the real dataset size of 424 gene trees, we determine an effective sample size of 143, roughly one-third 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 this recombination rate. We show the result in Fig 12; here we see that the relationship between ESS and N is approximately linear. Fitting a regression line through the origin gives the relationship

$$ESS = 0.314 \times N.$$

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

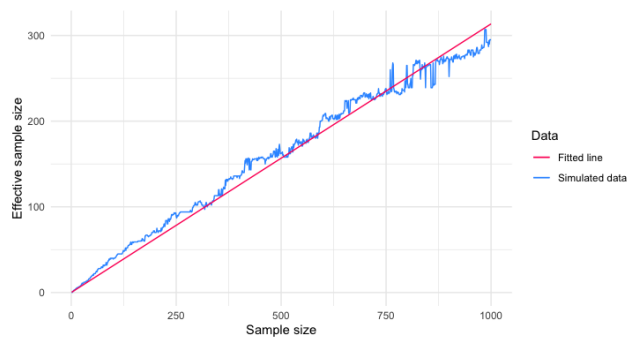

Fig 12. The effective sample size of ASTRAL under dependence, for fixed recombination rate $R = 1.26$.

Comparison with Wang and Liu

Wang and Liu [28] 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` [49]. 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 [50, 51].
2. LD100: As above, but with locus lengths of 100bp.
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).

We use a slightly different model to generate trees; our model is functionally equivalent to the sequentially Markov coalescent [36], a well-known approximation to

the coalescent with recombination. Although our model lacks the longer-range dependence structure of the true 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). Regardless, we do not expect the difference between these models to be significant.

The two major differences between our study and the TBIG/TBTG cases of [28] 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 [28], shown in Fig 21 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 $R = 0.1358$ (standard error 0.0002) for this tree.

We also repeated this process for the 37-taxon mammalian species tree above, finding an equivalent recombination rate of $R = 0.0183$ (standard error 0.0001), which is very low compared to our estimated recombination rate of 1.26. 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 [28].

We then fix the number of gene trees to the average produced by `ms` for the same parameters as [28]; 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 [52] model with $\alpha = 0.8$, as in [28], and then estimate gene trees with FastTree [50, 51]. Due to computational constraints, we run 20 replicates for inferred gene trees and 1000 replicates for true gene trees.

In Fig 13, 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.

We do note that the numerical accuracies we obtain are slightly different from the equivalent cases in [28]; 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.

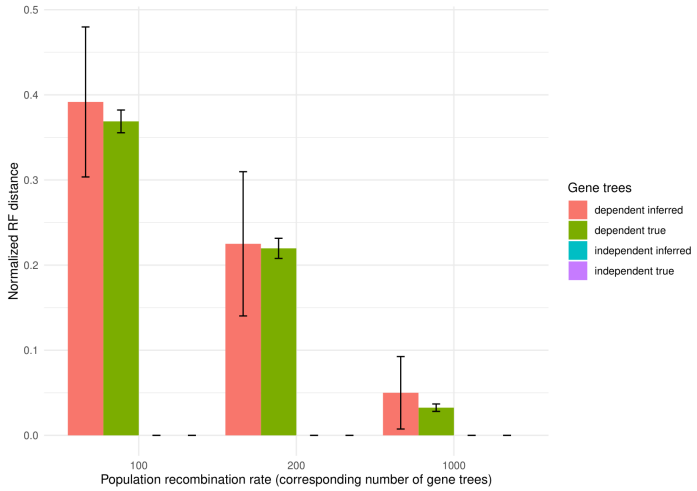

Fig 13. The accuracy of ASTRAL for the empirical tree of [28] for recombination rate $R = 0.1358$. Error bars represent standard error.

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 4.7, and the effective sample size by a factor of 3 for our 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, and indeed this is partially true for the relatively small number of taxa in the species tree we study here. However, for larger numbers of taxa, the required number of gene trees may well exceed the actual number of genes available. Furthermore, there is also a counter-effect: including more gene trees will lower the genomic distance between the trees, and thus increase the dependence between them, thus 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.

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 [26, 27], it has also been shown [28] 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 [53], which significantly complicates the problem.

To estimate a realistic recombination rate for the biological dataset, we used the average consecutive pairwise 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.

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.

Since we have shown gene tree dependence to have a detrimental effect on the accuracy of ASTRAL, this suggests that a new species tree inference method to take it 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.

References

1. Maddison WP. Gene trees and species trees. Systematic Biology. 1997;46(3):523–536.

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

3. Rannala B, Yang Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 2003;164(4):1645–1656.

4. Kingman JF. On the genealogy of large populations. Journal of Applied Probability. 1982;19(A):27–43.

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

6. Liu L, Yu L. Estimating species trees from unrooted gene trees. Systematic Biology. 2011;60(5):661–667.

7. Vachaspati P, Warnow T. ASTRID: accurate species trees from internode distances. BMC Genomics. 2015;16(10):1–13.

8. Sayyari E, Mirarab S. Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. BMC Genomics. 2016;17(10):101–113.

9. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 2014;30(17):i541–i548.

10. Mirarab S, Warnow T. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics. 2015;31(12):i44–i52.

11. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19(6):15–30.

12. Liu L, Yu L, Pearl DK, Edwards SV. Estimating species phylogenies using coalescence times among sequences. Systematic Biology. 2009;58(5):468–477.

13. Mossel E, Roch S. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Transactions on Computational Biology and Bioinformatics. 2008;7(1):166–171.

14. Giarla TC, Esselstyn JA. The challenges of resolving a rapid, recent radiation: empirical and simulated phylogenomics of Philippine shrews. Systematic Biology. 2015;64(5):727–740.

15. Ballesteros JA, Sharma PP. A critical appraisal of the placement of *Xiphosura* (*Chelicerata*) with account of known sources of phylogenetic error. Systematic Biology. 2019;68(6):896–917.

16. Yin J, Zhang C, Mirarab S. ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. Bioinformatics. 2019;35(20):3961–3969.

17. Davidson R, Vachaspati P, Mirarab S, Warnow T. Phylogenomic species tree estimation in the presence of incomplete lineage sorting and horizontal gene transfer. BMC Genomics. 2015;16(10):1–12.

18. Markin A, Eulenstein O. Quartet-based inference is statistically consistent under the unified duplication-loss-coalescence model. Bioinformatics. 2021;37(22):4064–4074.

19. Legried B, Molloy EK, Warnow T, Roch S. Polynomial-time statistical estimation of species trees under gene duplication and loss. Journal of Computational Biology. 2021;28(5):452–468.

20. Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Systematic Biology. 2022;71(2):367–381.

21. Molloy EK, Warnow T. To include or not to include: the impact of gene filtering on species tree estimation methods. Systematic Biology. 2018;67(2):285–303.

22. Patel S, Kimball RT, Braun EL. Error in phylogenetic estimation for bushes in the tree of life. Journal of Phylogenetics and Evolutionary Biology. 2013;1(2):1–10.

23. DeGiorgio M, Degnan JH. Robustness to divergence time underestimation when inferring species trees from estimated gene trees. Systematic Biology. 2014;63(1):66–82.

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

25. Huang H, Knowles LL. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Systematic Biology. 2016;65(3):357–365.

26. Lanier HC, Knowles LL. Is recombination a problem for species-tree analyses? Systematic Biology. 2012;61(4):691–701.

27. Zhu T, Flouri T, Yang Z. A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model. Molecular Ecology. 2022;31(10):2814–2829.

28. Wang Z, Liu KJ. A performance study of the impact of recombination on species tree analysis. BMC Genomics. 2016;17(10):165–174.

29. Conry M. Determining the impact of recombination on phylogenetic inference. The Florida State University; 2020.

30. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183–201.

31. Griffiths RC, Marjoram P. An ancestral recombination graph. Institute for Mathematics and its Applications. 1997;87:257.

32. Rasmussen MD, Hubisz MJ, Gronau I, Siepel A. Genome-wide inference of ancestral recombination graphs. PLOS Genetics. 2014;10(5):e1004342.

33. Mahmoudi A, Koskela J, Kelleher J, Chan Yb, Balding D. Bayesian inference of ancestral recombination graphs. PLOS Computational Biology. 2022;18(3):e1009960.

34. Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G. Inferring whole-genome histories in large population datasets. Nature Genetics. 2019;51(9):1330–1338.

35. Speidel L, Forest M, Shi S, Myers SR. A method for genome-wide genealogy estimation for thousands of samples. Nature Genetics. 2019;51(9):1321–1329.

36. McVean GA, Cardin NJ. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society B: Biological Sciences. 2005;360(1459):1387–1393.

37. Slatkin M, Pollack JL. The concordance of gene trees and species trees at two linked loci. Genetics. 2006;172(3):1979–1984.

38. Li Q, Scornavacca C, Galtier N, Chan Y. The multilocus multispecies coalescent: a flexible new model of gene family evolution. Systematic Biology. 2021;70(4):822–837.

39. Song S, Liu L, Edwards SV, Wu S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences. 2012;109(37):14942–14947.

40. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Mathematical Biosciences. 1981;53(1-2):131–147.

41. Spielman SJ, Wilke CO. Pyvolve: a flexible Python module for simulating sequences along phylogenies. PloS one. 2015;10(9):e0139047.

42. Tavaré S. Some probabilistic and statistical problems on the analysis of DNA sequence. Lecture of Mathematics for Life Science. 1986;17:57.

43. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-Likelihood phylogenies. Molecular Biology and Evolution. 2015;32(1):268–274.

44. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular Biology and Evolution. 2020;37(5):1530–1534.

45. Butler G, Rasmussen MD, Lin MF, Santos MA, Sakthikumar S, Munro CA, et al. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. Nature. 2009;459(7247):657–662.

46. Rasmussen MD, Kellis M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research. 2012;22(4):755–765.

47. Wu YC, Rasmussen MD, Bansal MS, Kellis M. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome Research. 2014;24(3):475–486.

48. Chan Y, Li Q, Scornavacca C. The large-sample asymptotic behaviour of quartet-based summary methods for species tree inference. Journal of Mathematical Biology. 2022;85(3):1–22.

49. Hudson RR. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics. 2002;18(2):337–338.

50. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Molecular Biology and Evolution. 2009;26(7):1641–1650.

51. Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. PLOS ONE. 2010;5(3):e9490.

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

53. Chan Y, Robin C. Reconciliation of a gene network and species tree. Journal of Theoretical Biology. 2019;472:54–66.

Supporting information

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		

Fig 14. Abbreviations of species names in the 37-taxon mammalian species tree in Fig 3.

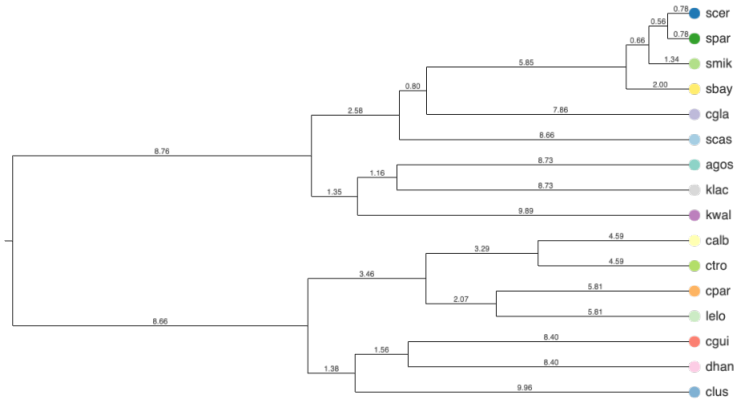

Fig 15. The 16-taxon fungi species tree with branch lengths in coalescent units. Species names are given in Table 16 in Supporting information.

Species name	Abbreviation
Candida alibicans	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

Fig 16. Abbreviations of species names in the 16-taxon fungi species tree in Fig 15 in Supporting information.

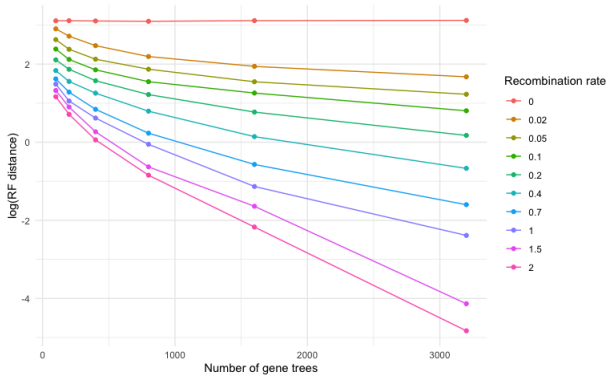

Fig 17. The accuracy of ASTRAL (log RF distance) vs the number of gene trees, averaged over 1000 replicates.

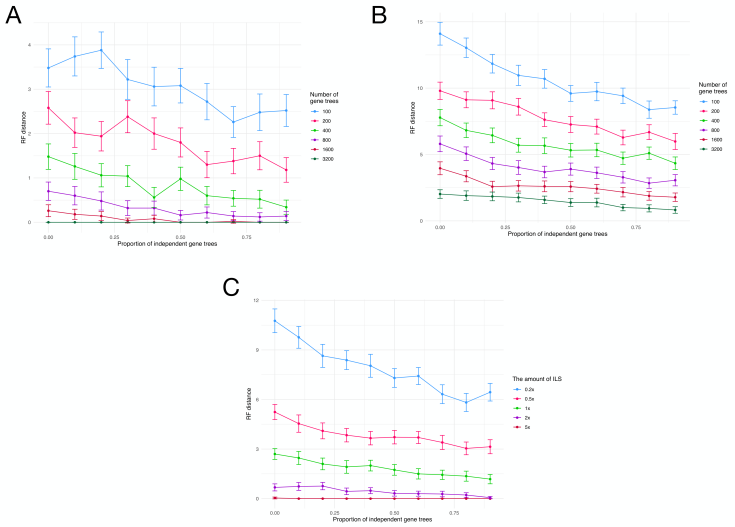

Fig 18. The accuracy of ASTRAL vs the proportion of independent gene trees for $R = 1.26$, averaged over 100 replicates. Error bars represent standard error. **A**: Default. **B**: Increased ILS (0.2×). **C**: Different ILS for $N = 200$.

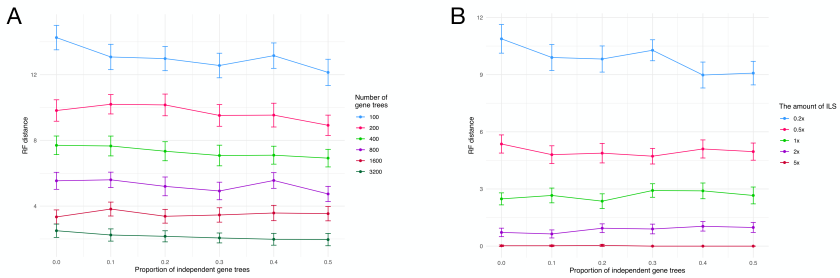

Fig 19. The accuracy of ASTRAL vs the proportion of independent gene trees, averaged over 100 replicates, with fixed overall recombination rate. Error bars represent standard error. The default settings are shown in Fig 10. **A**: Increased ILS (0.2×). **B**: Different ILS for $N = 200$.

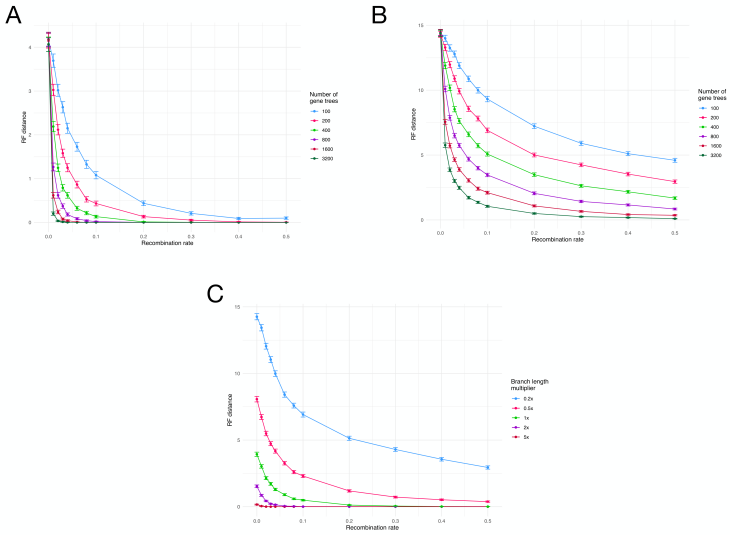

Fig 20. The accuracy of ASTRAL vs recombination rate for the 16-taxon fungi species tree, averaged over 1000 replicates. Error bars represent standard error. **A**: Default. **B**: Increased ILS (0.2×). **C**: Different ILS for $N = 200$.

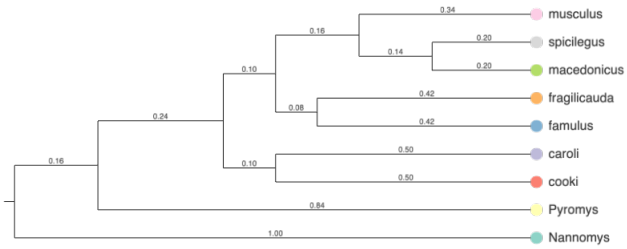

Fig 21. The 9-taxon mouse species tree from [28].