



Peer Community Journal

Section: Evolutionary Biology

Research article

Published
2026-03-24

Cite as

Wanting He, Celine Scornavacca and Yao-ban Chan (2026) *The effect of gene tree dependence on summary methods for species tree inference*, Peer Community Journal, 6: e25.

Correspondence

yaoban@unimelb.edu.au

Peer-review

Peer reviewed and recommended by PCI Evolutionary Biology, <https://doi.org/10.24072/pci.evolbiol.100860>



This article is licensed under the Creative Commons Attribution 4.0 License.

The effect of gene tree dependence on summary methods for species tree inference

Wanting He^{,1}, Celine Scornavacca^{,2}, and Yao-ban Chan^{,1}

Volume 6 (2026), article e25

<https://doi.org/10.24072/pcjournal.694>

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 accurate in many practical scenarios through extensive simulations. However, these simulations generally assume that the input gene trees are independent of each other (infinite recombination between loci). This is known to be unrealistic, as genes that are close to each other on the chromosome (or are co-evolving) have dependent phylogenies. 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 allow us to evaluate the impact of any level of dependence on the accuracy of ASTRAL, both when gene trees are known and estimated from sequences. We find that a fixed amount of dependence reduces the effective sample size by a constant factor. In current phylogenomic datasets, loci are generally sampled at large genomic distances to reduce gene tree dependence, thereby limiting the number of genes available for inference. However, full independence between genes is not required for accurate species tree estimation, and excluding gene trees may reduce inference accuracy. This creates a trade-off between the number of genes used and the degree of gene tree dependence. We therefore propose a method to identify the minimum genomic separation required to maintain satisfactory inference accuracy.

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

Peer Community Journal is a member of the
Centre Mersenne for Open Scientific Publishing
<http://www.centre-mersenne.org/>

e-ISSN 2804-3871



Introduction

Speciation is an evolutionary process where populations evolve and become distinct species (Coyne and Orr, 2004). A species tree, or phylogeny (Hillis et al., 1996), 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 (Darwin, 1859). 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 in 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.

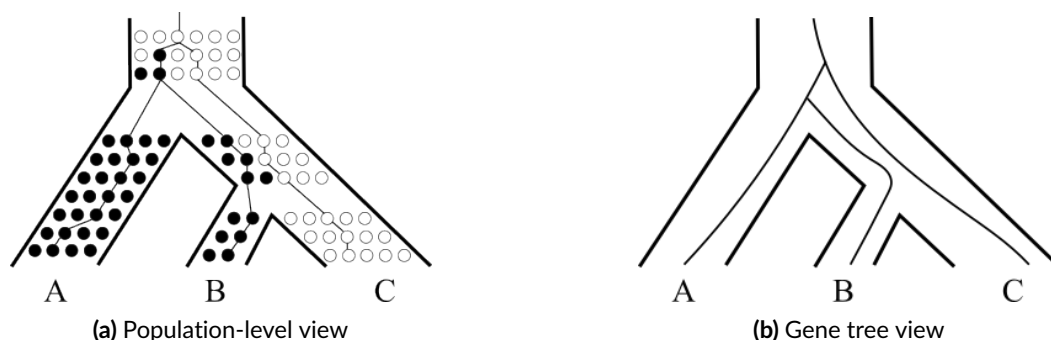


Figure 1 – An example of incomplete lineage sorting. Circles represent individuals of a population, and each row corresponds to a generation. Originally, the population contains only individuals with an ancestral allele (open circles), and then a mutation introduces a derived allele (filled circles), creating genetic polymorphism. The ancestral allele does not descend in species A, leaving only individuals with the derived allele. The polymorphism persists in the ancestor of species B and C. Subsequently species B retains only the derived allele, while species C retains only the ancestral 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. A common method used in the past was to concatenate sequences from multiple genes to form a supermatrix, and then obtain a species tree by applying traditional phylogenetic inference methods, such as distance-based or maximum-likelihood. This paradigm implicitly assumes that all genes come from 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., 2014b; 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 (Minh et al., 2020; Nguyen et al., 2015; Stamatakis, 2014), Bayesian (Flouri et al., 2018; Heled and Drummond, 2009; Ogilvie et al., 2017; Rannala and Yang, 2017), and co-estimation methods (Heled and Drummond, 2009; Liu, 2008; Ogilvie et al., 2017).

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 theoretically to give a statistically consistent estimator of the species tree if input gene trees are sampled under the MSC model (Mirarab et al., 2014b), 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, showing that ASTRAL is highly accurate under the MSC model (Mirarab et al., 2014b; 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 in simulations, they were almost all 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. However, these studies explored only a limited range of simulation scenarios, and more complex cases involving short internal branches, high substitution rate variation, and deep divergences among taxa remain largely untested.

In this paper, we focus on the latter assumption; in reality, gene trees are dependent because genes can be 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 sampling gene trees from an independent model, 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 whole 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 statistical 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, 1996). 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.

Phylogenomic data can be produced in many different ways. A common approach is to identify orthologous gene families between the species of interest and extract the gene sequences. This results in full sequences that have some amount of genomic distance between them, as typically only exons are extracted. The dependence between the genes depends on the amount of separation; ideally the distance is sufficient to achieve approximately independent genes, but this is not always checked and the conditions for achieving a sufficient distance are not well studied. An alternative approach is to analyse the SNPs common to the sequences and separate them into loci in some way, often using linkage disequilibrium-based (LD) approaches to ensure approximately independent loci. As SNPs are evenly distributed among the genome, this approach discards some data to ensure sufficient separation.

Our aim in this paper is to reassess the accuracy of ASTRAL under conditions when gene trees are dependent. 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. This allows us to numerically estimate the impact of a number of factors, including the number of loci, the amount of ILS, recombination rate, and the dependence structure, on the accuracy of ASTRAL. For real data, we study approaches to ensuring that gene dependence does not affect the accuracy of species tree inference. We re-analyse an 8-taxon mouse SNP dataset (Liu et al., 2015), and find that satisfactory accuracy can be achieved by sampling SNPs at distances much shorter than those suggested by LD-based methods. For datasets produced by extracting orthologous genes, we devise a method to determine the minimum genomic distance in order to maintain satisfactory accuracy. The real-life 37-taxon mammalian dataset that we study (Mirarab et al., 2014b) has genes separated by much larger distances than this minimum, indicating that current datasets are minimally affected by gene tree dependence.

Methods

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 use it to produce 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).

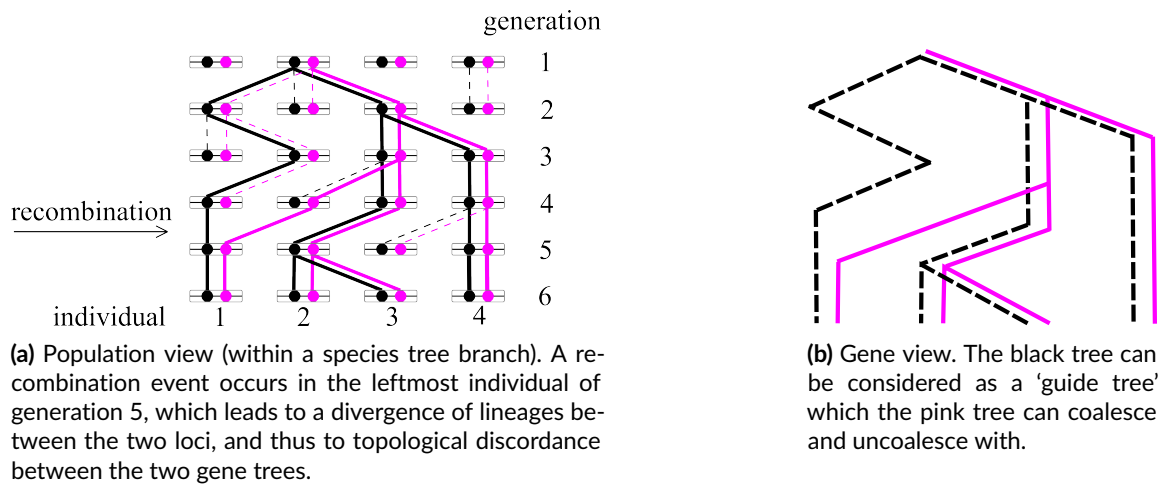


Figure 2 – An example of the two-locus model. The black and pink lines represent the genealogies in the first and second loci respectively within a population (a species branch), and the black tree is considered as the guide tree.

Consider first a single panmictic population. 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. (If there is a genomic distance of d sites between loci, then in population genetics notation, $R = 2N_e rd$, where $2N_e$ 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 that lineage 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 two 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.

This allows us to generate a gene tree conditional on another gene tree. To generate multiple dependent gene trees along a linear genome, we first generate a gene tree under the MSC. Each gene tree is then used as a guide tree in our model for the subsequent gene tree. For a linear genome, this reduces to the sequentially Markov coalescent (SMC; McVean and Cardin, 2005), sampled at constant intervals along the genome. However, our model is more flexible, as it can also simulate dependent gene trees where there is a non-linear dependence structure between genes (for example, when dependence is produced by functional rather than proximity relationships). As far as we know, our algorithm is the only one that can achieve this, although we do not use this functionality in this paper.

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 (Chen et al., 2009), 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 as described above retains more flexibility. Nevertheless, we also conduct some simulations on ultrametric trees to determine that the results from our model do not differ significantly from those produced using gene trees generated by `msprime` (see [Comparison with 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 use 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.

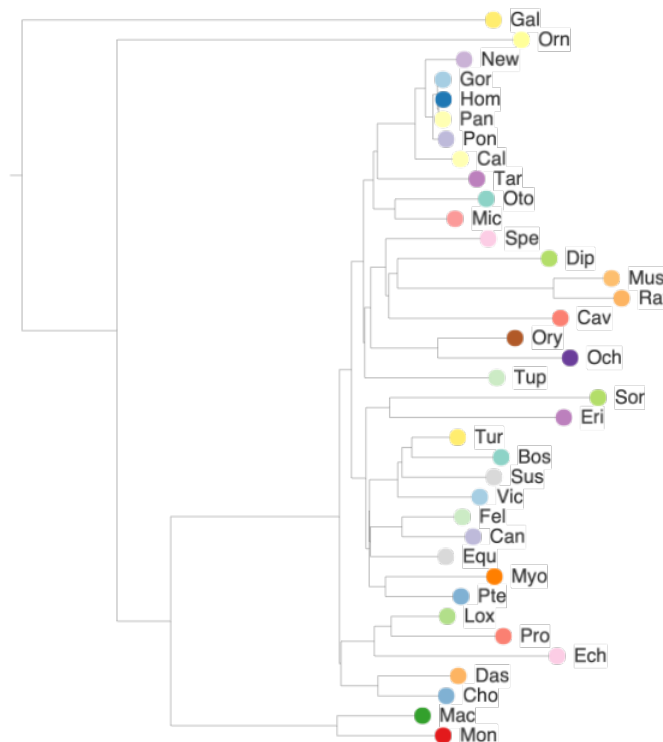
Simulations

We explore the performance of ASTRAL using simulations 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., 2014b). The species tree 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 3099bp. We re-estimate the branch lengths as specified in [Supporting information](#), while keeping the same topology. The result is shown in Figure 3.

We perform simulations with the same parameter settings as in Mirarab et al. (2014b), adding dependence between the gene trees as specified above. We then use these trees as input to ASTRAL to infer a species tree, and evaluate the inference 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 counting the number of clades present in one tree but not the other.

Additionally, we rescale the simulated dependent gene trees into units of substitutions per site, and then evolve gene sequences along the trees with Pyvolve (Spielman and Wilke, 2015), under the GTR+ model of site evolution (Tavaré, 1986; Yang, 1994) with parameters estimated from the biological dataset (see Supporting information for more details). As in Mirarab et al. (2014b), we use three sequence lengths: 500bp, 1000bp, and a mixture where half of the sequences are 500bp and the other half are 1000bp. 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. Although this range is much smaller than the values observed in the dataset (see Section Estimating minimum genomic distance in a phylogenomic dataset for details), we nevertheless focus on it, as it captures the regime in which accuracy is most sensitive to changes in R . 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.



5.10

Figure 3 – The 37-taxon mammalian species tree (with branch lengths in coalescent units) (full species names are given in Table S2 in Supporting information).

It has been shown that ASTRAL performs worse with an increased amount of ILS (Mirarab et al., 2014b), so we are also 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. (2014b).

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 $N = 800$ 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 0.1.

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, 3200]	$0.2 \times$ 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 .

To verify that our results are not specific to the 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 in Figure S6 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. 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. Although the 37-taxon mammalian species tree we study is not ultrametric, we extend the terminal branches to obtain an ultrametric tree, and compare our model with `msprime`.

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) 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.
- (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.

- 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. <https://doi.org/10.1109/TCBB.2008.66>.
- 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. <https://doi.org/10.1093/molbev/msu300>.
- Ogilvie HA, Bouckaert RR, Drummond AJ (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular biology and evolution* **34**, 2101–2114. <https://doi.org/10.1093/molbev/msx126>.
- Pamilo P, Nei M (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution* **5**, 568–583. <https://doi.org/10.1093/oxfordjournals.molbev.a040517>.
- Patané JS, Martins Jr J, Setubal JC (2024). A Guide to Phylogenomic Inference. *Comparative Genomics: Methods and Protocols* **267**–345. https://doi.org/10.1007/978-1-0716-3838-5_11.
- 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**–10. <https://doi.org/10.4172/2329-9002.1000110>.
- Phifer-Rixey M, Harr B, Hey J (2020). Further resolution of the house mouse (*Mus musculus*) phylogeny by integration over isolation-with-migration histories. *BMC evolutionary biology* **20**, 120. <https://doi.org/10.1186/s12862-020-01666-9>.
- 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. <https://doi.org/10.1093/molbev/msp077>.
- Price MN, Dehal PS, Arkin AP (2010). FastTree 2 approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**, e9490. <https://doi.org/10.1371/journal.pone.0009490>.
- 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. <https://doi.org/10.1093/genetics/164.4.1645>.
- Rannala B, Yang Z (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic biology* **66**, 823–842. <https://doi.org/10.1093/sysbio/syw119>.
- Rasmussen MD, Kellis M (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research* **22**, 755–765. <https://doi.org/10.1101/gr.123901.111>.
- Robinson DF, Foulds LR (1981). Comparison of phylogenetic trees. *Mathematical Biosciences* **53**, 131–147. [https://doi.org/10.1016/0025-5564\(81\)90043-2](https://doi.org/10.1016/0025-5564(81)90043-2).
- Sayyari E, Mirarab S (2016). Anchoring quartet-based phylogenetic distances and applications to species tree reconstruction. *BMC Genomics* **17**, 101–113. <https://doi.org/10.1186/s12864-016-3098-z>.
- Simmons MP, Sloan DB, Gatesy J (2016). The effects of subsampling gene trees on coalescent methods applied to ancient divergences. *Molecular Phylogenetics and Evolution* **97**, 76–89. <https://doi.org/10.1016/j.ympev.2015.12.013>.
- Slatkin M, Pollack JL (2006). The concordance of gene trees and species trees at two linked loci. *Genetics* **172**, 1979–1984. <https://doi.org/10.1534/genetics.105.049593>.

- 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. <https://doi.org/10.1073/pnas.1211733109>.
- Spielman SJ, Wilke CO (2015). Pyvolve: a flexible Python module for simulating sequences along phylogenies. *PLoS one* **10**, e0139047. <https://doi.org/10.1371/journal.pone.0139047>.
- Stamatakis A (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
- Taillon-Miller P, Bauer-Sardiña I, Saccone NL, Putzel J, Laitinen T, Cao A, Kere J, Pilia G, Rice JP, Kwok PY (2000). Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature genetics* **25**, 324–328. <https://doi.org/10.1038/77100>.
- Tavaré S (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Some Mathematical Questions in Biology: DNA Sequence Analysis*, 17. Lectures on Mathematics in the Life Sciences. American Mathematical Society, pp. 57–86.
- Vachaspati P, Warnow T (2015). ASTRID: accurate species trees from internode distances. *BMC Genomics* **16**, 1–13. <https://doi.org/10.1186/1471-2164-16-S10-S3>.
- Wang Z, Liu KJ (2016). A performance study of the impact of recombination on species tree analysis. *BMC Genomics* **17**, 165–174. <https://doi.org/10.1186/s12864-016-3104-5>.
- 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. <https://doi.org/10.1101/gr.161968.113>.
- 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 paralogy. *Systematic Biology* **71**, 367–381. <https://doi.org/10.1093/sysbio/syab056>.
- Yang Z (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution* **39**, 306–314. <https://doi.org/10.1007/BF00160154>.
- Yin J, Zhang C, Mirarab S (2019). ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics* **35**, 3961–3969. <https://doi.org/10.1093/bioinformatics/btz211>.
- Zhang C, Mirarab S (2022a). ASTRAL-Pro 2: ultrafast species tree reconstruction from multi-copy gene family trees. *Bioinformatics* **38**, 4949–4950. <https://doi.org/10.1093/bioinformatics/btac620>.
- Zhang C, Mirarab S (2022b). Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. *Molecular Biology and Evolution* **39**, msac215. <https://doi.org/10.1093/molbev/msac215>.
- 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. <https://doi.org/10.1186/s12859-018-2129-y>.
- Zhang C, Scornavacca C, Molloy EK, Mirarab S (2020). ASTRAL-Pro: quartet-based species-tree inference despite paralogy. *Molecular Biology and Evolution* **37**, 3292–3307. <https://doi.org/10.1093/molbev/msaa139>.

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. <https://doi.org/10.1111/mec.16433>.