



3.4:2 The Concatenation Question

Chapters 1.2 and 1.4 [Stamatakis and Kozlov 2020; Lartillot 2020]), and improved computers, made combined analysis with heterogeneous models tractable, placating both camps.

Heterogeneity in the substitution process is one thing; heterogeneity in tree topology is another. The fact that different loci can have different underlying histories has been recognised for a long time (Hudson, 1983; Tajima, 1983; Pamilo and Nei, 1988). However, it was not until coalescent theory leaked from population genetics into systematics (Maddison, 1997) that many realised how ubiquitous gene tree discordance could be. We can now be confident that much of the topological variation in recent divergences we see among loci is due to biological causes rather than technical errors (e.g. Brawand et al., 2014; *Heliconius* Genome Consortium, 2012; Fontaine et al., 2015; Novikova et al., 2016; Pease et al., 2016; Pollard et al., 2006; Rogers et al., 2019).

The heightened awareness of gene tree discordance appears to have revived the total evidence *versus* consensus debate, albeit in a different guise. Those on the (new) consensus side, armed with the multispecies coalescent, argue that gene trees should be inferred separately and then combined to infer a species tree (e.g. Edwards, 2009). Those on the (new) total evidence side pick holes in these arguments and remind readers of older arguments in support of concatenation (e.g. Gatesy and Springer, 2014).

If history repeats itself again, the revived debate will die a natural death. New, improved model-based methodologies will eventually address the concerns of both camps. In the interim, the debate motivates the discussion of fundamental issues related to the scale and scope of gene tree discordance, the importance of statistical consistency, and relative sources of error in phylogenetic inference. It also provides a convenient framework for a chapter discussing those issues.

In Section 2 we introduce much of the relevant theory about the population processes that lead to discordance, discussing incomplete lineage sorting, expected branch lengths, the “anomaly zone”, and the impact of recombination. In Section 3 we discuss *summary tree methods*, approaches that estimate a gene tree for each locus and then combine these estimates. We discuss the consistency, and inconsistency, of these methods, and suggest that chopping the genome up into small chunks for separate analyses might open up these methods to systematic error and biases.

In Section 4 we examine the approach of concatenating all genes together before analysis, effectively ignoring the potential discordance among trees. This wholesale concatenation has been proven to be statistically inconsistent (Roch and Steel, 2015), though, as we point out, the branch lengths used in the proof are ridiculously short. The question of whether or not to concatenate is a question of finding a compromise between different kinds of error. The error from discordance exists for both recent and deep divergences, though with deep divergences it appears that alternative sources of error become much more important.

Section 5 examines one of the most confusing threads in the debate: whether or not we should tolerate recombination within loci. Oddly, both sides accuse the other of ignoring recombination and discordance. We discuss the arguments for and against combining trees, and some of the ways that have been proposed to overcome the bias associated with short loci. We also **consider** the pitfalls when considering clustered or binned genes as single loci in the multispecies coalescent.

In the final section we examine alternatives to summary methods and concatenated maximum likelihood, while also noting that there are excellent reasons for estimating individual gene trees, **independent** of species tree inference.

2 Gene tree discordance and the multispecies coalescent

One of the most important findings from genome-scale data in phylogenomics is that gene tree discordance is ubiquitous. Recognizing discordance between gene trees—and accounting for it in the inference of species trees—has been a major focus of the last decade of phylogenetic methods development. Among all of the possible causes of discordance, incomplete lineage sorting (ILS) has received the most attention, though introgression between species may well have a comparable real impact (Mallet et al., 2016). Here we focus on ILS, a concept we introduce in Section 2.2. While gene duplication and subsequent loss is also often included as a biological cause of discordance (e.g. Degnan and Rosenberg, 2009; Maddison, 1997), it is due to the mis-assignment of paralogs as orthologs (see Chapter 2.4 [Fernández et al. 2020]), and we do not consider it further.

2.1 Basic coalescent thinking

Two randomly chosen sequences at a locus from a single population share a common ancestor in the recent past. Under the Wright-Fisher model of diploid, hermaphroditic organisms with effective population size N , the probability that two autosomal sequences sampled from a single generation find a common ancestor in the previous generation is $1/2N$. We use $2N$ here because each of the N individuals carries two copies of this locus; alternatively, we can imagine a population of haploid individuals of size $2N$. A simple outcome of Mendelian inheritance is that the distribution of times back until two lineages find a common ancestor—that is, until they “coalesce”—is exponentially distributed with a mean of $2N$ generations. This process has a large variance, and independent loci sampled from the same two individuals will coalesce at many different times in the past.

Results for samples of size $n > 2$ can be derived under the n -coalescent model (Hudson, 1983; Kingman, 1982; Tajima, 1983). With $n = 3$ there are three equally probable topologies relating three lineages within a single population, and with $n = 4$ there are 18 equally probable labeled histories (by “labeled history” we mean that we distinguish between trees with the same relationships but that have lineages coalescing in a different temporal order). For all such topologies the coalescent model provides expectations for the times to coalescence, which in turn also imply branch lengths upon which mutations can accumulate (see Hein et al., 2004 and Wakeley, 2009 for an overview).

Importantly, in the n -coalescent model we assume that all observed mutations are neutral. This assumption allows us to completely separate the genealogical process of coalescence from the process by which mutations occur in the sample history. Every locus in this model has an underlying gene tree, irrespective of whether we are able to determine what it is from the pattern of informative mutations—our ability to infer a tree is not a necessary condition for its existence. More complex coalescent models than those described here are available, some incorporating selection, and some with non-Wright-Fisher populations (e.g. Spence et al., 2016). The importance of such models for phylogenetics is a largely unexplored area.

2.2 Incomplete lineage sorting

The small but finite amount of time it takes lineages to coalesce has significant consequences for variation in gene tree topologies. One useful way to think about this phenomenon is to ask whether all of the sampled lineages in a population have found their common ancestor before some pre-specified time in the past. Avise et al. (1983) referred to the case where all lineages find their common ancestor as “lineage sorting”. Conversely, we now refer to

3.4:4 The Concatenation Question

the case in which there are two or more lineages remaining as “incomplete lineage sorting” (ILS).

To be concrete, consider time measured in coalescent units, so that $T = t/2N$, where t is the number of generations. Given exponentially distributed coalescence times, the probability of lineage sorting of two lineages by time T in the past (i.e. the probability of 2 lineages going to 1 lineage) is:

$$P_{21}(T) = 1 - e^{-T}. \quad (1)$$

Likewise, the probability of incomplete lineage sorting (i.e. the probability of 2 lineages staying as 2 lineages) is:

$$P_{22}(T) = e^{-T}. \quad (2)$$

This result implies that only $\approx 63\%$ of loci will have coalesced by the mean expected time to coalescence ($2N$ generations, or 1 coalescent time unit), but that 95% of loci will have coalesced by $6N$ generations in the past. If we consider a species with an effective population size of 100,000 and a generation time of 1 year, these numbers imply that it would take on average 600,000 years for 95% of loci to sort. Similar calculations for the probability of lineage sorting among more than two lineages can also be made (Tavaré, 1984).

Incomplete lineage sorting is important for phylogenetics because it implies that, even when two species share an ancestral branch, not every gene tree sampled from those species will also have that branch. Enumerating the probabilities of specific topologies and their associated branch lengths in the presence of ILS is the goal of the *multispecies coalescent* model, which we discuss next.

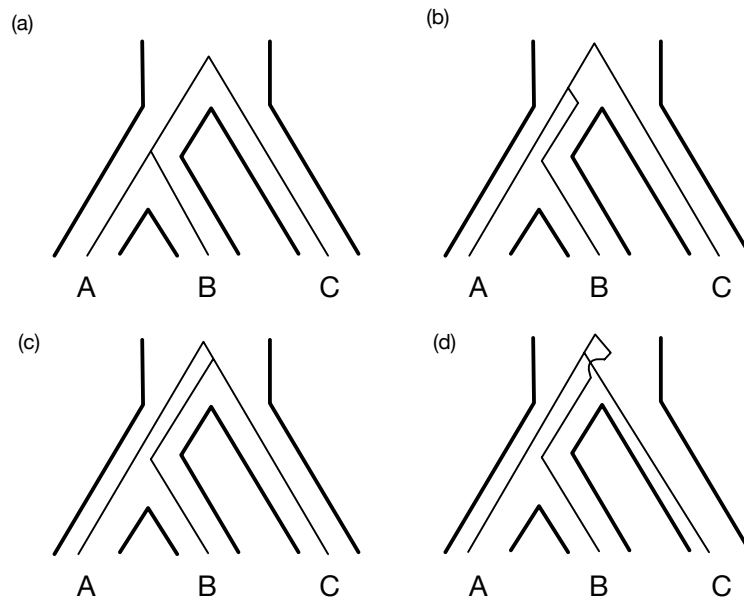
2.3 The frequency of different topologies under the multispecies coalescent

Because of the time required for lineage sorting, ancestral populations that existed between closely spaced speciation events become very important. The multispecies coalescent (MSC) model (Hudson, 1983; Pamilo and Nei, 1988; Tajima, 1983; Takahata and Nei, 1985; Takahata, 1989) recognizes that coalescence in ancestral populations can determine the frequency and branch lengths of different topologies, and attempts to quantify these measures. The MSC is limited in many ways—it does not include many processes that can be modeled in the general coalescent framework—but it does provide an important guide to the effects of ILS on gene tree discordance.

Imagine that we have sampled one (haploid) individual from each of three species, A , B , and C , and that the true relationship between the species is $((A, B), C)$ (see Figure 1). We would like to know the probability of sampling a gene tree that matches this topology if we collect data from a single locus. Discordance at a locus can occur if the lineages sampled from A and B do not coalesce in their common ancestral population, and instead one of them coalesces with the lineage from C in the population ancestral to all three species. Regardless of how long the tip branches are (because no coalescence between species can occur along them), the probability that A and B do not coalesce in the most recent shared ancestral population is given by Equation 2, with T denoting the length of this internal branch. If there is no coalescence (i.e. if ILS occurs) then each of the three possible topologies are equally likely to occur in the common ancestral population of A , B , and C .

Under this model, the expected frequencies of the two discordant topologies are both (Hudson, 1983):

$$E[f_{((A,C),B)}] = E[f_{((B,C),A)}] = (1/3)e^{-T}. \quad (3)$$



■ **Figure 1** Incomplete lineage sorting and gene tree discordance. (a) Complete lineage sorting, so that the gene tree is consistent with the species tree. (b)-(d) Incomplete lineage sorting, of which only the first gives a gene tree consistent with the species tree.

A concordant topology is also produced under ILS, at the same frequency as the two discordant topologies (Hudson, 1983).

A concordant topology must be produced if there is lineage sorting (with probability $1 - e^{-T}$), so the total frequency of concordant topologies is:

$$E[f_{((A,B),C)}] = (1/3)e^{-T} + (1 - e^{-T}) = 1 - (2/3)e^{-T}. \quad (4)$$

We can see that there is more discordance with very small internal branch lengths (up to a maximum of $2/3$ of all trees), and that at very long internal branch lengths there is essentially no discordance due to ILS. Following from the example given above, at $T = 6$ approximately 95% of loci will have sorted in the common ancestor of A and B , and will therefore be concordant. Of the remaining 5%, $1/3$ will also be concordant, with the other loci equally split between the two discordant topologies.

Similar calculations can be made for arbitrarily large numbers of lineages undergoing ILS (Degnan and Salter, 2005). With four taxa undergoing ILS, there are now two internal branches of any species tree that must be considered, with ILS occurring in either one or both branches. While there are 18 possible labeled histories with four taxa, often only the 15 unlabeled histories are considered (e.g. Rosenberg, 2002), as we do not distinguish between, for instance, the two different possible sequences of coalescences in the topology $((A, B), (C, D))$ (either (A, B) first or (C, D) first). The number of possible topologies quickly explodes with more taxa. It is essential to realize, however, that these calculations reflect the number of lineages undergoing ILS, not the number of taxa in a tree. It may be that even in a tree of 100 taxa only 3 lineages are in a phylogenetic “knot” that induces ILS. In such cases we need only concern ourselves with ILS calculations for three taxa.

One of the most important take-home messages about the MSC is that ILS can occur at any time in the past. As can be seen from Equations 3 and 4, the only parameter determining the probability of discordance due to ILS is T , which measures the length of an

3.4:6 The Concatenation Question

internal branch of the species tree in coalescent units (though note that most species trees are reported using absolute time or numbers of mutations per site per branch). Whether this internal branch existed 1 million or 100 million years ago, the amount of discordance due to ILS will be the same. However, our ability to determine a gene tree topology, and to ascribe it to ILS or not, *is* certainly dependent on how long ago these events occurred and how long the internal branches of individual gene trees are.

2.4 Gene tree branch lengths under the MSC

The expected branch lengths in both concordant and discordant gene trees are easily obtainable from the MSC model. As the coalescent process can have no effect on tip branch lengths after the most recent speciation event when one sequence is sampled from each species, total branch lengths will always have tip branch lengths added as a constant. Therefore, we focus on the expected lengths of gene tree branches above the tips.

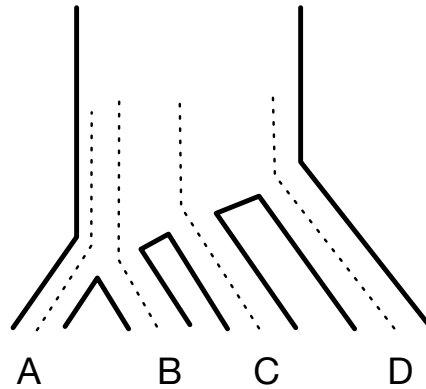
When ILS occurs among three lineages, two coalescent events occur in the common ancestral population. Looking backwards, the first coalescence is expected to occur $2N/3$ generations in the past (Figure 1(b-d)). Following this event, which is equally likely to join any of the three possible pairs of sequences, there are two lineages left in the tree, and therefore an average of $2N$ more generations until the entire sample has reached its most recent common ancestor. This result means that the internal branch of any topology that is the result of ILS (whether concordant or discordant) has an expected length of $2N$ generations. Even in “hard polytomies”, where the length of the internal branch of the species tree is zero, each of the three gene trees has an internal branch with expected length $2N$. In contrast, the internal branch of loci that complete the lineage sorting process have a minimum expected length of $2N$ generations and a maximum length equal to the length of the internal branch of the species tree. Mendes and Hahn (2018) provide analytical formulas for these expectations.

To give some perspective on how long such an internal branch is, recall that the expected number of pairwise differences between two sequences within a population is $4N\mu$, where μ is the per-nucleotide mutation rate (often this compound parameter is referred to as θ). The expected number of mutations on the internal branch of discordant trees is $2N\mu$, or about half the number of pairwise differences. As the proportion of sites with such differences are generally at or below 1% for multicellular organisms (Leier et al. 2012), we can begin to understand why it is so hard to accurately identify discordant gene trees. Note that this internal branch of discordant trees has the same short length no matter how far back in time the ILS occurred and no matter how long the internal branch of the species tree is. This is why discordant gene trees due to ILS will always have lower bootstrap support than any concordant tree—because they always have a shorter internal branch—and why using a bootstrap cut-off to determine which gene trees to include in an analysis could result in a biased estimate of discordance.

2.5 The anomalous anomaly zone

Under the multispecies coalescent, each species tree determines a distribution on the set of gene trees. There have been several theoretical results on this distribution, mainly for the case where exactly one individual is sampled from each species. It is tempting to think of this distribution as a cloud of random gene trees, centred on the species tree. Degnan and Rosenberg (2006) showed that this picture can be misleading. One can construct a species tree such that the most likely gene tree under the multispecies coalescent is *not* the species

tree. This observation, perhaps more than any other, has been used to justify methods that account for incomplete lineage sorting in inferring the species tree.



■ **Figure 2** Tree used to generate an anomaly zone in Degnan and Rosenberg (2006). The branches below the root are short enough that almost all coalescent events occur in the population at the root. As the coalescent process in one population results in more trees that are balanced (e.g. $((A, B), (C, D))$), these gene trees will have higher probability under the MSC than the underlying species tree.

Consider the species tree in Figure 2. If the two successive internal branches in this species tree are short enough (in coalescent units), coalescences are most likely to occur in the ancestral population of all four lineages. Under the coalescent model in a single population, symmetric trees, like $((A, B), (C, D))$ have higher probability than asymmetric (“caterpillar”) trees like $((A, B), C), D$ because the former are associated with two labeled histories while the latter are associated with only a single labeled history. As a result, if the species tree is a caterpillar and most coalescence occurs in the single ancestral population, an asymmetric gene tree matching the species tree can be less common than one of the symmetric gene trees. The term “anomaly zone” is used to describe the area of parameter space (in terms of branch lengths in the species tree), where the gene tree concordant with the species tree is less probable than some other tree.

Despite the anomaly zone’s bogeyman-like status in phylogenetics, it is not as scary as it looks. For instance, there is no anomaly zone for gene trees with branch lengths. The times between coalescent events have an exponential density. The density function of an exponential random variable is strictly decreasing, so gets larger for values close to zero. The mode, or most “likely” value, for an exponential random variable is zero, and the most likely time of coalescence is zero or effectively instantaneous. As a consequence, the mode of the distribution of gene trees with branch lengths under the MSC is identical to the species tree! For the same reason, in the multispecies coalescent with extremely small population sizes (such that pairs of lineages coalesce as soon as possible), every gene tree would match the species tree exactly.

Nevertheless the anomaly zone *does* cause problems with methods for inferring species trees based on counts or frequencies of gene trees. It was also thought that the anomaly zone was responsible for consistency problems with concatenated maximum likelihood, though as we will see later it is not the anomaly zone that is responsible.

2.6 The coalescent with recombination

The classical MSC model makes two important assumptions about the role of recombination, neither of which is likely to be true in real data but both of which are required to produce the topological distributions expected under the MSC. The first assumption is that we are dealing with individually non-recombining loci, such that each locus or gene contains only a single underlying topology. Non-recombining loci such as mtDNA, cpDNA, or the Y (or W) chromosome all conform to this assumption. For sequences drawn from the autosomal nuclear genome, the length of non-recombining loci is a function of rates of recombination and population sizes (N).

These considerations raise the question of how long we expect non-recombining loci in the nuclear genome to be. The rate of recombination varies along the genome and across species. In humans, the average length of non-recombining autosomal loci is 4.8-5.9 kilobases (International HapMap Consortium, 2005), though there is a huge variance in the length of such blocks. For species with larger population sizes such as *Drosophila*, there is more effective recombination and block sizes are commensurately smaller, on the order of hundreds of bases or less (Hey and Nielsen, 2004). However, because the amount of nucleotide diversity also scales with population size, the large block sizes in species with small populations like humans do not necessarily result in more phylogenetic resolution within each locus.

While a strict interpretation of the MSC assumes non-recombining loci, there are some kinds of recombinations that have no effect on inference. If the recombination is limited to lineages within a branch, different sites will have identical gene trees, even if they are undergoing recombination.

The real concern of intra-locus recombination for inference is when there are multiple histories present among loci, as when there is incomplete lineage sorting or introgression. When this occurs, different sites within a single protein-coding gene can have discordant gene trees. In fact, Mendes et al. (2019) showed that 70% and 91% of protein-coding genes in primates and *Drosophila*, respectively, contain two or more gene tree topologies from a single phylogenetic knot (i.e. three species undergoing ILS). Even if genes were on average the same length as non-recombining stretches of chromosome, multiple trees will be combined unless the recombination events exactly flank the sequence being used for inference. When there are multiple knots across a larger tree, the length of loci that do not have a single recombination event within them at any point in the tree can become vanishingly small (Gatesy and Springer, 2014). We return to the issue of intra-locus recombination below, as it affects all of the methods we discuss here.

There is a second critical assumption that the MSC makes about recombination: that different loci have independent gene trees (conditional on the species tree). In other words, it assumes that there is sufficient recombination between loci that gene trees for different loci are independent (conditional on the species tree). Non-independence of samples is a common problem across statistics, known to cause greater variability than expected (overdispersion). The consequences of assuming independence in phylogenetics are not well understood, but generally assumptions of this type result in greater confidence in results than is warranted.

3 Summary gene tree methods

In a summary tree method, separate gene trees are estimated for each locus, and these gene trees are then used to infer the species tree. In the contemporary revival of the total evidence *versus* consensus debate, those advocating summary tree methods fall squarely in the “consensus” camp. Examples of this approach include ASTRAL (Mirarab et al.,

In summary, then, there is no question that ILS is an important cause of gene tree discordance, especially when looking at recently diverged populations or species. It is not clear, however, that the phylogenetic “error” due to ILS trumps all other sources of error, especially as we move into the more distant past. It is not that problems due to ILS get smaller, only that the error due to all other causes gets much larger.

5 The role of recombination in the debate

5.1 Concatalescence

One issue that we have mostly avoided discussing so far is whether the loci analyzed by summary tree methods are themselves non-recombining. A standard unit of phylogenetic analysis is the protein-coding gene. As mentioned earlier, the vast majority of protein-coding genes in eukaryotic genomes are likely to contain two or more topologies in the presence of ILS. While single exons less often contain multiple topologies (Mendes et al., 2019), they are much shorter and are therefore both less likely to be able to fully resolve trees containing many taxa and will provide many fewer sites with which to fit complex substitution models. The implicit compromise of using even single protein-coding genes is that we have enough sequence with which to carry out “good enough” phylogenetic analyses, even though we may be violating the MSC. This compromise approach has been given the portmanteau “concatalescence” (Gatesy and Springer, 2014).

There has been quite a kerfuffle in the literature surrounding concatalescence (Gatesy and Springer, 2014; Liu et al., 2014b; Springer and Gatesy, 2016; Edwards et al., 2016). It seems to us that the main question is not whether current approaches using single protein-coding genes violate the MSC—they almost certainly do—but what effect this violation has on the inferred topologies, and especially the distribution of inferred topologies.

We have already seen that in extreme cases of ILS, concatenated ML will converge on the wrong tree with more and more data (Kubatko and Degnan, 2007). What is less clear is the behavior of shorter genes that combine only a handful of different topologies. Some simulations have been done to examine the effect of recombination on realistic gene lengths (Lanier and Knowles, 2012), finding little effect relative to other sources of phylogenetic error. However, these simulations have been criticized as having too little recombination (because the intervening introns were not taken into account; Gatesy and Springer, 2014), and did not seem to include the areas of parameter space where concatenated ML fails. If individual gene trees are biased by concatenation, then so too will be the rooted triplets and unrooted quartets extracted from them for use with summary methods.

Regardless of the criticisms of published simulations, researchers in favor of summary gene tree methods face an apparent paradox: if typical protein-coding genes are immune to the effects of recombination, ILS, and concatenation (and can therefore be used to construct gene trees), then why not concatenate all the loci? Unfortunately, no theory yet exists that bounds the amount of recombination and ILS allowable while still producing correct trees. Further work is clearly needed to know how far such methods can be pushed.

5.2 Conditional concatenation and binning

Summary methods are problematic because ML is biased. We have now seen two causes for this bias: sequences that are too short to accurately model the substitution process (section 3.3) and sequences that are so long that they contain multiple conflicting topologies within them (section 5.1). In the next section we discuss “full” likelihood methods that can possibly

with recombination.

The most widely used methods can be conveniently separated into two groups: those that use blocks of sequence, but assume no recombination within loci and free recombination between loci, and those that use only variable sites, but assume that there is free recombination between them. In the first category are methods implemented in BPP (Rannala and Yang, 2017), StarBeast (Heled and Drummond, 2009; Ogilvie et al., 2017) and PhyloNet (Wen et al., 2018). All three methods can work directly from individual gene alignments, calculating the likelihood of the data under the MSC. They accommodate sampling error in the gene trees that summary tree methods ignore. These methods (or their extensions; Zhang et al., 2018) are also able to infer species networks—essentially the species tree with reticulations—though the methods that do so require time-consuming MCMC sampling. Regardless of the way in which tree space is explored, these methods still assume that each input gene is a non-recombining unit, and therefore face some of the same modelling questions as summary methods.

In the second category are methods that assume free recombination between individually varying sites. Methods that use this type of data are varied, including SNAPP (Bryant et al., 2012), PoMo (De Maio et al., 2015), and SVDquartets (Chifman and Kubatko, 2014). While SNAPP and PoMo are optimal for species with recent splits (and multiple individuals sampled per species), SVDquartets is able to infer species trees with deep splits. These methods all avoid the issues with short non-recombining blocks of sequence, completely circumventing the problem by combining together a large number of independently evolving loci. Although complex substitution models incorporating all the different kinds of rate variation observed are not yet included in the tools listed here, these methods are some of the most promising for the future of phylogenetics.

6.2 Why genes should still be analysed separately

Even if you believe that species tree inference should only be carried out with concatenated data, it is still useful to infer trees for each gene. Arguments in favor of the examination of individual gene trees go back as far as the consensus/total-evidence debate (e.g. De Queiroz, 1993), and genomic data has only made this more true. Individual gene trees can reveal an enormous amount about variation in history along the genome, different rates of evolution in different genomic compartments, and different potential biases or patterns in a dataset. The signal in the data is in the variable history among loci, not just species relationships (Bravo et al., 2019).

One obvious example of where the study of individual gene trees can help is in cases of horizontal gene transfer (HGT) or introgression between species. The disagreement among trees is widely considered to be the best evidence for transfer (Soucy et al., 2015). Similarly, gene flow between sexually reproducing species can result in gene tree discordance at introgressed loci. The distribution of discordant trees along the genome is one of the few indications that introgression is occurring (e.g. Liu et al., 2014a), and the distinct heights and branch lengths of introgressed trees can help to disentangle complex histories (e.g. Fontaine et al., 2015; Kearns et al., 2018).

Because genes underlie traits, gene trees may also be a much better guide to trait evolution than species trees, especially when there is a lot of discordance (Hahn and Nakhleh, 2016). In cases with extreme levels of discordance, such as adaptive radiations, it may even be possible to associate individual discordant loci with incongruent traits (e.g. Pease et al., 2016; Wu et al., 2018). Radiations may be one of the best arguments for approaches that examine individual gene trees, as it becomes highly unlikely that *any* locus follows the inferred

species history (e.g. Jarvis et al., 2014).

Finally, the visualization of gene tree heterogeneity may itself be a worthwhile endeavor. Hillis et al. (2005) showed how a collection of inferred gene trees could be visualized in tree space using multidimensional scaling. Duchêne et al. (2017) used this multidimensional scaling approach to identify clusters of gene tree topologies supporting conflicting resolutions of the species tree, and were able to show that the clusters were generated by ILS. Other sorts of visualization tools may be equally useful in different contexts (e.g. Esser et al., 2004).

6.3 Moving forward

Phylogenetic inference is a hard problem, especially for deep divergences. As we have seen, much of the difficulty stems from how and what to model, and the extent to which different models impact on our inference.

Therefore, the choice of methods to use should be informed by the largest sources of error. At shallower timescales gene trees can be accurately inferred and ILS (and introgression) can be large sources of variance among gene trees. At deeper timescales the sources of variance flip, such that ILS becomes relatively less important. ILS certainly occurs at deep timescales, but many other processes also come into play, making the inference of individual gene trees much harder. While we hope that researchers interested in resolving relationships at, for instance, the base of animals keep the possibility of gene tree discordance in mind, it is certainly understandable that the methods they employ to infer a species tree do not model this process explicitly.

References

- Avise, J. C., Shapira, J. F., Daniel, S. W., Aquadro, C. F., and Lansman, R. A. (1983). Mitochondrial DNA differentiation during the speciation process in *Peromyscus*. *Molecular Biology and Evolution*, 1:38–56.
- Bayzid, M. S., Mirarab, S., Boussau, B., and Warnow, T. (2015). Weighted statistical binning: enabling statistically consistent genome-scale phylogenetic analyses. *PLoS One*, 10(6):e0129183.
- Bravo, G. A., Antonelli, A., Bacon, C. D., Bartoszek, K., Blom, M. P. K., Huynh, S., Jones, G., Knowles, L. L., Lamichhaney, S., Marcussen, T., Morlon, H., Nakhleh, L. K., Oxelman, B., Pfeil, B., Schliep, A., Wahlberg, N., Werneck, F. P., Wiedenhoeft, J., Willows-Munro, S., and Edwards, S. V. (2019). Embracing heterogeneity: coalescing the tree of life and the future of phylogenomics. *PeerJ*, 7:e6399.
- Brawand, D., Wagner, C. E., Li, Y. I., Malinsky, M., Keller, I., Fan, S., Simakov, O., Ng, A. Y., Lim, Z. W., Bezault, E., Turner-Maier, J., Johnson, J., Alcazar, R., Noh, H. J., Russell, P., Aken, B., Alfoldi, J., Amemiya, C., Azzouzi, N., Baroiller, J.-F., Barloy-Hubler, F., Berlin, A., Bloomquist, R., Carleton, K. L., Conte, M. A., D’Cotta, H., Eshel, O., Gagné, L., Galibert, F., Gante, H. F., Gnerre, S., Greuter, L., Guyon, R., Haddad, N. S., Haerty, W., Harris, R. M., Hofmann, H. A., Hourlier, T., Hulata, G., Jaffe, D. B., Lara, M., Lee, A. P., MacCallum, I., Mwaiko, S., Nikaido, M., Nishihara, H., Ozouf-Costaz, C., Penman, D. J., Przybylski, D., Rakotomanga, M., Renn, S. C. P., Ribeiro, F. J., Ron, M., Salzburger, W., Sanchez-Pulido, L., Santos, M. E., Searle, S., Sharpe, T., Swofford, R., Tan, F. J., Williams, L., Young, S., Yin, S., Okada, N., Kocher, T. D., Miska, E. A., Lander, E. S., Venkatesh, B., Fernald, R. D., Meyer, A., Ponting, C. P.,

- Streelman, J. T., Lindblad-Toh, K., Seehausen, O., and Di Palma, F. (2014). The genomic substrate for adaptive radiation in african cichlid fish. *Nature*, 513(7518):375–381.
- Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., and RoyChoudhury, A. (2012). Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, 29(8):1917–1932.
- Bull, J. J., Huelsenbeck, J. P., Cunningham, C. W., Swofford, D. L., and Waddell, P. J. (1993). Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42(3):384–397.
- Chifman, J. and Kubatko, L. (2014). Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324.
- De Maio, N., Schrempf, D., and Kosiol, C. (2015). PoMo: an allele frequency-based approach for species tree estimation. *Systematic Biology*, 64(6):1018–1031.
- De Queiroz, A. (1993). For consensus (sometimes). *Systematic Biology*, 42(3):368–372.
- Degnan, J. H. (2013). Anomalous unrooted gene trees. *Systematic Biology*, 62(4):574–590.
- Degnan, J. H. and Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, 2(5):e68.
- Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution*, 24(6):332–340.
- Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution*, 59(1):24–37.
- Duchêne, D. A., Bragg, J. G., Duchêne, S., Neaves, L. E., Potter, S., Moritz, C., Johnson, R. N., Ho, S. Y. W., and Eldridge, M. D. B. (2017). Analysis of phylogenomic tree space resolves relationships among marsupial families. *Systematic Biology*, 67(3):400–412.
- Edwards, S. V. (2009). Is a new and general theory of molecular systematics emerging? *Evolution*, 63(1):1–19.
- Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S., Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., and Davis, C. C. (2016). Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular Phylogenetics and Evolution*, 94:447–462.
- Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., Henze, K., Kretschmann, E., Richly, E., Leister, D., Bryant, D., Steel, M. A., Lockhart, P. J., Penny, D., and Martin, W. (2004). A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular Biology and Evolution*, 21(9):1643–1660.
- Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4):401–410.
- Fernández, R., Gabaldón, T., and Dessimoz, C. (2020). Orthology: Definitions, prediction, and impact on species phylogeny inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.4, pages 2.4:1–2.4:14. No commercial publisher | Authors open access book.
- Fontaine, M. C., Pease, J. B., Steele, A., Waterhouse, R. M., Neafsey, D. E., Sharakhov, I. V., Jiang, X., Hall, A. B., Catteruccia, F., Kakani, E., Mitchell, S. N., Wu, Y.-C., Smith, H. A., Love, R. R., Lawniczak, M. K., Slotman, M. A., Emrich, S. J., Hahn, M. W., and Besansky, N. J. (2015). Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, 347(6217):1258524.
- Gatesy, J. and Springer, M. S. (2014). Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Molecular Phylogenetics and Evolution*, 80:231–266.

- Gillespie, J. H. and Langley, C. H. (1979). Are evolutionary rates really variable? *Journal of Molecular Evolution*, 13:27–34.
- Hahn, M. W. and Nakhleh, L. (2016). Irrational exuberance for resolved species trees. *Evolution*, 70:7–17.
- Hein, J., Schierup, M., and Wiuf, C. (2004). *Gene genealogies, variation and evolution: a primer in coalescent theory*. Oxford University Press, USA.
- Heled, J. and Drummond, A. J. (2009). Bayesian Inference of Species Trees from Multilocus Data. *Molecular Biology and Evolution*, 27(3):570–580.
- Heliconius Genome Consortium (2012). Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, 487:94–98.
- Hey, J. and Nielsen, R. (2004). Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, 167(2):747–760.
- Hillis, D. M., Heath, T. A., and St John, K. (2005). Analysis and visualization of tree space. *Systematic Biology*, 54(3):471–482.
- Hudson, R. R. (1983). Testing the constant-rate neutral allele model with protein-sequence data. *Evolution*, 37(1):203–217.
- Inagaki, Y., Susko, E., Fast, N. M., and Roger, A. J. (2004). Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaeobacteria in EF-1 α phylogenies. *Molecular Biology and Evolution*, 21(7):1340–1349.
- International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299.
- Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S., Gabaldón, T., Capella-Gutiérrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M., Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li, N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V., Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdociimi, F., Samaniego, J. A., Velazquez, A. M. V., Alfaro-Núñez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield, P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng, Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J., Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson, W., Koepfli, K.-P., O’Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R., Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alström, P., Edwards, S. V., Stamatakis, A., Mindell, D. P., Cracraft, J., et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.
- Je roy, O., Brinkmann, H., Delsuc, F., and Philippe, H. (2006). Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22(4):225–231.
- Kearns, A. M., Restani, M., Szabo, I., Schröder-Nielsen, A., Kim, J. A., Richardson, H. M., Marzlu, J. M., Fleischer, R. C., Johnsen, A., and Omland, K. E. (2018). Genomic evidence of speciation reversal in ravens. *Nature Communications*, 9(1):906.
- Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, 13(3):235–248.
- Kubatko, L. S. and Degnan, J. H. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology*, 56(1):17–24.

- Kück, P., Mayer, C., Wägele, J.-W., and Misof, B. (2012). Long branch effects distort maximum likelihood phylogenies in simulations despite selection of the correct model. *PLoS ONE*, 7(5):e36593.
- Lanier, H. C. and Knowles, L. L. (2012). Is recombination a problem for species-tree analyses? *Systematic Biology*, 61:691–701.
- Lartillot, N. (2020). The bayesian approach to molecular phylogeny. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.4, pages 1.4:1–1.4:17. No commercial publisher | Authors open access book.
- Leigh, J. W., Lapointe, F.-J., Lopez, P., and Baptiste, E. (2011). Evaluating phylogenetic congruence in the post-genomic era. *Genome Biology and Evolution*, 3:571–587.
- Liu, K. J., Dai, J., Truong, K., Song, Y., Kohn, M. H., and Nakhleh, L. (2014a). An HMM-based comparative genomic framework for detecting introgression in eukaryotes. *PLoS Computational Biology*, 10:e1003649.
- Liu, L. and Edwards, S. V. (2009). Phylogenetic analysis in the anomaly zone. *Systematic Biology*, 58:452–460.
- Liu, L., Xi, Z., and Davis, C. C. (2014b). Coalescent methods are robust to the simultaneous effects of long branches and incomplete lineage sorting. *Molecular Biology and Evolution*, 32(3):791–805.
- Liu, L., Yu, L., and Edwards, S. V. (2010). A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302.
- Maddison, W. P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3):523–536.
- Mallet, J., Besansky, N., and Hahn, M. W. (2016). How reticulated are species? *BioEssays*, 38(2):140–149.
- Martin, D. P., Lemey, P., Lott, M., Moulton, V., Posada, D., and Lefevre, P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*, 26(19):2462–2463.
- Mendes, F. K. and Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Systematic Biology*, 67(1):158–169.
- Mendes, F. K., Livera, A. P., and Hahn, M. W. (2019). The perils of intralocus recombination for inferences of molecular convergence. *Philosophical Transactions of the Royal Society B*, 374:20180244.
- Mirarab, S., Bayzid, M. S., and Warnow, T. (2016). Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Systematic Biology*, 65(3):366–80.
- Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S., and Warnow, T. (2014). ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics*, 30(17):i541–i548.
- Novikova, P. Y., Hohmann, N., Nizhynska, V., Tsuchimatsu, T., Ali, J., Muir, G., Guggisberg, A., Paape, T., Schmid, K., Fedorenko, O. M., Holm, S., Sall, T., Schlotterer, C., Marhold, K., Widmer, A., Sese, J., Shimizu, K. K., Weigel, D., Kramer, U., Koch, M. A., and Nordborg, M. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nature Genetics*, 48(9):1077–1082.
- Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, 34(8):2101–2114.
- Page, R. D. (1996). On consensus, confidence, and “total evidence”. *Cladistics*, 12(1):83–92.

3.4:22 REFERENCES

- Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583.
- Pease, J., Haak, D., Hahn, M. W., and Moyle, L. C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biology*, 14:e1002379.
- Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., and Baurain, D. (2011). Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biology*, 9(3):e1000602.
- Philippe, H. and Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biology*, 9(1):91.
- Philippe, H., Vienne, D. M. d., Ranwez, V., Roure, B., Baurain, D., and Delsuc, F. (2017). Pitfalls in supermatrix phylogenomics. *European Journal of Taxonomy*, 283:1–25.
- Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N., and Delsuc, F. (2005). Heterotachy and long-branch attraction in phylogenetics. *BMC Evolutionary Biology*, 5(1):50.
- Pollard, D. A., Iyer, V. N., Moses, A. M., and Eisen, M. B. (2006). Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. *PLoS Genetics*, 2(10):e173.
- Rannala, B., Edwards, S. V., Leaché, A., and Yang, Z. (2020). The multi-species coalescent model and species tree inference. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 3.3, pages 3.3:1–3.3:21. No commercial publisher | Authors open access book.
- Rannala, B. and Yang, Z. (2017). Efficient Bayesian species tree inference under the multispecies coalescent. *Systematic Biology*, page syw119.
- Richards, E. J., Brown, J. M., Barley, A. J., Chong, R. A., and Thomson, R. C. (2018). Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Systematic biology*, 67(5):847–860.
- Roch, S., Nute, M., and Warnow, T. (2018). Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. *Systematic Biology*, 68(2):281–297.
- Roch, S. and Steel, M. (2015). Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theoretical Population Biology*, 100:56–62.
- Rogers, J., Muthuswamy, R., Harris, R. A., Mailund, T., Leppälä, K., Athanasiadis, G., Schierup, M. H., Cheng, J., Munch, K., Walker, J. A., Konkel, M. K., Jordan, V., Steely, C. J., Beckstrom, T. O., Bergey, C., Burrell, A., Schrepf, D., Noll, A., Kothe, M., Kopp, G. H., Liu, Y., Murali, S., Billis, K., Martin, F. J., Mu ato, M., Cox, L., Else, J., Disotell, T., Muzny, D. M., Phillips-Conroy, J., Aken, B., Eichler, E. E., Marques-Bonet, T., Kosiol, C., Batzer, M. A., Hahn, M. W., Tung, J., Zinner, D., Roos, C., Jolly, C. J., Gibbs, R. A., Worley, K. C., and the Baboon Genome Analysis Consortium (2019). The comparative genomics and complex population history of *Papio* baboons. *Science Advances*, 5:eaau6947.
- Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology*, 61(2):225–247.
- Scornavacca, C. and Galtier, N. (2017). Incomplete lineage sorting in mammalian phylogenomics. *Systematic Biology*, 66:112–120.
- Simion, P., Delsuc, F., and Philippe, H. (2020). To what extent current limits of phylogenomics can be overcome? In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 2.1, pages 2.1:1–2.1:34. No commercial publisher | Authors open access book.

- Soucy, S. M., Huang, J., and Gogarten, J. P. (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16:472–482.
- Spence, J. P., Kamm, J. A., and Song, Y. S. (2016). The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561.
- Springer, M. S. and Gatesy, J. (2016). The gene tree delusion. *Molecular Phylogenetics and Evolution*, 94:1–33.
- Stamatakis, A. and Kozlov, A. M. (2020). Efficient maximum likelihood tree building methods. In Scornavacca, C., Delsuc, F., and Galtier, N., editors, *Phylogenetics in the Genomic Era*, chapter 1.2, pages 1.2:1–1.2:18. No commercial publisher | Authors open access book.
- Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O., and Rogers, J. S. (2001). Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic Biology*, 50(4):525–539.
- Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460.
- Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics*, 122(4):957–966.
- Takahata, N. and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, 110(2):325–344.
- Tavaré, S. (1984). Line-of-descent and genealogical processes, and their applications in population genetics models. *Theoretical Population Biology*, 26(2):119–164.
- Vachaspati, P. and Warnow, T. (2015). ASTRID: Accurate species trees from internode distances. *BMC Genomics*, 16(10):S3.
- Wakeley, J. (2009). Coalescent theory. *Roberts & Company, Greenwood Village, Colorado*.
- Wang, H.-C., Susko, E., and Roger, A. J. (2019). The relative importance of modeling site pattern heterogeneity versus partition-wise heterotachy in phylogenomic inference. *Systematic Biology*, 68(6):1003–1019.
- Warnow, T. (2015). Concatenation analyses in the presence of incomplete lineage sorting. *PLoS Currents*, page doi: 10.1371/currents.tol.8d41ac0f13d1abedf4c4a59f5d17b1f7.
- Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. (2018). Inferring phylogenetic networks using PhyloNet. *Systematic Biology*, 67(4):735–740.
- White, D. J., Bryant, D., and Gemmell, N. J. (2013). How good are indirect tests at detecting recombination in human mtDNA? *G3: Genes, Genomes, Genetics*, 3(7):1095–1104.
- Wu, M., Kostyun, J. L., Hahn, M. W., and Moyle, L. C. (2018). Dissecting the basis of novel trait evolution in a radiation with widespread phylogenetic discordance. *Molecular Ecology*, 27:3301–3316.
- Zhang, C., Ogilvie, H. A., Drummond, A. J., and Stadler, T. (2018). Bayesian inference of species networks from multilocus sequence data. *Molecular Biology and Evolution*, 35(2):504–517.