Molecular Phylogenetics and Evolution 181 (2023) 107724

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Comparing inference under the multispecies coalescent with and without recombination

Zhi Yan *, Huw A. Ogilvie, Luay Nakhleh

Department of Computer Science, Rice University, 6100 Main Street, Houston 77005, TX, USA

ARTICLE INFO

Keyword:
Multispecies coalescent
Recombination
Species tree
Divergence time
Population size

ABSTRACT

Accurate inference of population parameters plays a pivotal role in unravelling evolutionary histories. While recombination has been universally accepted as a fundamental process in the evolution of sexually reproducing organisms, it remains challenging to model it exactly. Thus, existing coalescent-based approaches make different assumptions or approximations to facilitate phylogenetic inference, which can potentially bring about biases in estimates of evolutionary parameters when recombination is present. In this article, we evaluate the performance of population parameter estimation using three methods—StarBEAST2, SNAPP, and diCal2—that represent three different types of inference. We performed whole-genome simulations in which recombination rates, mutation rates, and levels of incomplete lineage sorting were varied. We show that StarBEAST2 using short or medium-sized loci is robust to realistic rates of recombination, which is in agreement with previous studies. SNAPP, as expected, is generally unaffected by recombination events. Most surprisingly, diCal2, a method that is designed to explicitly account for recombination, performs considerably worse than other methods under comparison.

1. Introduction

The development of statistical models has moved phylogenetics beyond the qualitative estimates of cladistic relationships to inference of demographic histories including divergence times, population size changes and other quantitative values. These statistical methods rely on the fact that different loci across the genome have distinct histories which arise from recombination during sexual reproduction.

Instances of discordance have long been documented between the evolutionary history of a set of species and the evolutionary histories of individual loci within the genomes of those species. While such discordance patterns arise due to biological processes acting within and across species, their analysis is confounded in practice by systematic error in the data. One biological source of discordance that is ubiquitous across the Tree of Life is incomplete lineage sorting, or ILS. When an ancestral population splits into two descendant populations, ancestral polymorphism could be maintained across multiple speciation events leading to conflict between individual gene genealogies and species phylogenies—a phenomenon known as ILS. To account for ILS in species tree inference, the multispecies coalescent (MSC) has emerged as the dominant model as it naturally accounts for coalescent stochasticity (Rannala and Yang, 2003). The model was later extended to operate on phylogenetic networks, giving rise to the multispecies network coalescent (MSNC), in order to account for reticulate evolutionary histories in the presence of ILS (Yu et al., 2014).

Species phylogeny inference methods under the MSC (and MSNC) can be classified into two categories. Summary, or gene-tree-based, methods use gene tree estimates as the input data and infer species phylogenies that summarize these gene trees under various criteria. While these gene-tree methods are usually fast and scalable, their accuracy is adversely affected by gene tree estimation error (Patel et al., 2013). The second category consists of methods that directly infer the species phylogeny from molecular sequence data. While methods in the second category are often more accurate and produce more information on the evolutionary parameters, they are generally less scalable than gene-tree-based methods (Zimmermann et al., 2014).

Most implementations of species phylogeny inference under the MSC assume that their input is structured into non–recombining loci or coalescence genes (Rannala and Yang, 2003; Heled and Drummond, 2009; Bryant et al., 2012a; Ogilvie et al., 2017). These “c-genes” (Doyle, 1997; Springer and Gatesy, 2016) correspond to the segments between recombination events in a genome, and therefore each site within a c-gene will share the same phylogenetic history as every other site in the same c-gene. However, in practice these implementations are instead

* Corresponding author.
E-mail addresses: zhi.yan@rice.edu (Z. Yan), Huw.A.Ogilvie@rice.edu (H.A. Ogilvie), nakhleh@rice.edu (L. Nakhleh).

https://doi.org/10.1016/j.ympev.2023.107724
Received 28 August 2022; Received in revised form 17 January 2023; Accepted 24 January 2023
Available online 28 January 2023

used to analyze data sets structured into m-genes, "a particular sequence of nucleotides along a molecule of DNA…which represents a functional unit of inheritance" (Rieger et al., 2012; Doyle, 2021). Because this is common practice, we seek to assess the impact of recombination within m-genes on the inference of the divergence times and population sizes by these methods when m-genes are used as input. We will refer to these implementations using "multilocus methods" as shorthand.

There are a few studies to date which have investigated the impact of recombination on species phylogeny estimation. Lanier and Knowles (2012) and Wang and Liu (2016) assessed the performance of coalescent-aware methods in the presence of recombination and have shown that species tree topology inference under the MSC appears to be robust to intra-locus recombination. In examining the impact of violations of free inter-locus recombination, Wang and Liu (2016) additionally found that the use of recombination breakpoints for identifying loci improves the accuracy of species tree topology estimation. However, little is known about the extent to which recombination affects the estimation of ancestral population sizes and divergence times. Very recently, Zhu et al. (2022) tested the effects of intra-locus recombination on several inference problems including the estimation of population parameters under the MSC. They found that in "realistic" simulations, Bayesian methods using multilocus sequence data under the MSC performed reasonably well when the amount of recombination was low or moderate, but could underestimate population sizes and inflate species divergence time given elevated recombination rates, confirming similar observations from previous studies (Wall, 2003; Lohse and Frantz, 2014).

The estimation of divergence times along with population sizes of species phylogenies can be conducted in a variety of ways and we choose three modern and representative approaches to study. First, the multilocus method StarBEAST2 (Ogilvie et al., 2017) jointly infers gene and species histories and is an extensively used approach to infer evolutionary parameters while accounting for rate variation and gene discordance. Second, the single nucleotide polymorphism (SNP) method SNAPP (Bryant et al., 2012b) avoids the problem of c-gene/m-gene conflation by assuming the input data set consists of unlinked biallelic markers such as SNPs, since recombination cannot occur within a single site. Third, advances in whole-genome sequencing have been driving the development of another class of methods that use sequentially Markovian approximations of the coalescent to integrate over gene histories and recombination breakpoints (Steinrücken et al., 2019; Liu et al., 2021), thus conducting inference under the (multispecies) coalescent with recombination, and we used diCal2 (Steinrücken et al., 2019) to represent this class.

While some degree of model violation is unavoidable for any statistical method when applied to real data, we show that adding recombination and linkage has a similar and mild effect on StarBEAST2 and SNAPP. However, parameters inferred by diCal2 could be wildly erroneous, casting doubt on the approximations employed by that method. Our results add more support to the use of methods that assume recombination-free loci, even when the assumption is violated.

2. Materials and methods

2.1. Simulations

To examine the impact of recombination on continuous parameter estimation given a fixed species topology, we performed whole-genome simulations under a classical coalescent with recombination model using msprime version 1.0.2 (Kelleher et al., 2016). We varied the species divergence times, mutation rates, and recombination rates to create data sets approximating different evolutionary scenarios, which were broadly similar to those estimated from modern humans and other hominids (Roach et al., 2010). Specifically, two model species trees with shallow and deep evolutionary timescales were used: i) a shallow phylogeny of height 1.5 million years, and ii) a deep phylogeny of height 18 million years (Fig. 1). We considered two per-generation mutation rates, $\mu = 10^{-8}$/site/generation and $\mu = 10^{-7}$/site/generation, where the smaller rate is very similar to the modern human rate of 1.1×10^{-8}/site/generation reported by Roach et al. (2010). We used three different recombination rates: $r = 2 \times 10^{-9}$, $r = 2 \times 10^{-8}$, and $r = 2 \times 10^{-7}$ per site per generation, where the medium rate is similar to the human genome-wide average recombination rate of 1.26 cM/Mb reported in Jensen-Seaman et al. (2004). For each combination of μ and r, 10 replicates of ten 1-Mb long chromosomal segments were simulated for each taxon. In all cases, the population size was 20,000 diploid individuals, and the generation time was 25 years as in Li and Durbin (2011). DNA sequences were generated assuming the Jukes-Cantor nucleotide substitution model.

2.2. Continuous-parameter estimation

For the inference of divergence times and population sizes, the true species tree topology (((A,B),C),D) (Fig. 1) was provided to all methods evaluated here.

The Markov chain Monte Carlo (MCMC) convergence of StarBEAST2 and SNAPP was examined by the Gelman-Rubin convergence diagnostic $\widehat{R}$ (Gelman and Rubin, 1992), and the effective sample size (ESS).

2.2.1. StarBEAST2

To prepare the sequence alignments for StarBEAST2, five alignments of length 200, 1000, and 5000 bp were extracted from each chromosomal segment by evenly sampling across the segments using python script from https:// github.com/ForBioPhylogenomics/tutorials/blob/main/week2_src/make_alignments_from_vcf.py.

The StarBEAST2 analyses used the Jukes-Cantor substitution model with a strict clock where the clock rate was set to the true mutation rate, and assumed a Yule prior on the species tree where Uniform(0,0.01) was used as prior distribution on the speciation rate. Constant population sizes were estimated given a $1/X$ hyperprior (popMean).

The chain length was set to 0.5 billion generations, sampling every 200,000 states, and the first 20 % of collected samples were discarded as burn-in. Three independent runs (using different seeds) of such chains were carried out on every data set. We generated maximum clade credibility (MCC) trees with posterior mean heights using TreeAnnotator v2.6.6 (Drummond et al., 2012).

2.2.2. SNAPP

For each data set, the segments from 10 chromosomes were concatenated into one file via BCFtools's concat command (Danecek

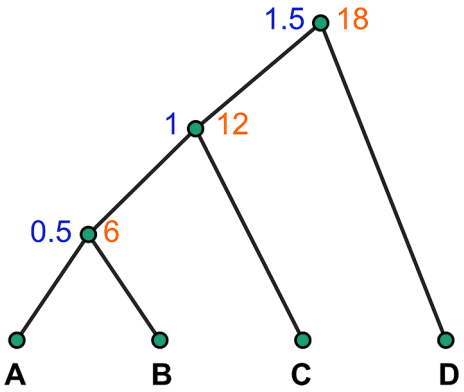

Fig. 1. The model species tree used in the simulations. Values in blue and orange at internal nodes represent the node times in million years for the shallow and deep phylogenies, respectively. Population size of 20,000 individuals and generation time of 25 years were assumed. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

et al., 2021). The SNAPP input XML files were produced by the ruby script snapp_prep.rb (Stange et al., 2018). Specifically, the ruby script extracted at most 1000 SNPs, where the minimum distance between SNPs was 1000 nucleotides, with only variable sites included. We used a $1/X$ prior on the speciation rate lambda in the Yule model. We placed a gamma-distributed prior on the population size parameter theta, with an alpha (shape parameter of the gamma distribution) value of 0.5 and beta (rate parameter of the gamma distribution) values of 33 and 3, for the low- and high-mutation-rate scenarios, respectively. The clock rate was fixed to the true mutation rate. To enable ascertainment correction, we included the number of constant sites that is ascertained for.

MCMC chains were run for 24 h of wall-clock time with a single thread (on average, each chain was run over 40 million generations) and sampled every 5,000 generations. The first 30 % samples of each chain were discarded as burn-in. Three independent chains were run on each replicate data set. The post-burn-in samples were summarized as an MCC tree with nodes scaled to the mean height estimates using TreeAnnotator v2.6.6 (Drummond et al., 2012).

2.2.3. diCal2

diCal2 was given as input the simulated haplotypes, the true recombination rate, and the true mutation rate. For each generation, we used 70 particles, each of which performed 6 EM-steps, and 6 best points were selected as the parents for the next generation. Time was discretized into intervals by 11 break points, which was chosen log-uniformly between 1000 years and 300 million years. The genetic algorithm was repeated for 5 generations as used in Steinrücken et al. (2019). The parameters with highest composite likelihoods were reported.

3. Results

3.1. Characteristics of the simulation data

Because the accumulation of recombination is a result of the product of interaction between recombination rate and time span, deep divergence tends to yield shorter c-genes compared with the recent divergence, with a mean length slightly over 62 % of that of the shallow phylogeny (Table 1). At both tree depths, increases in recombination rate result in shorter c-genes (Table 1). Based on our simulations, we observed that an order of magnitude increase in the recombination rate leads to roughly an order of magnitude decrease in the average c-gene size. As expected, the mutation rate has negligible impact on the lengths of c-genes (results are now shown in the table).

3.2. Accuracy of continuous parameter estimates

Overall, all methods inferred better estimates of divergence times compared with population sizes. Bayesian inference based on the MSC using either linked sites (i.e., StarBEAST2) or SNPs (i.e., SNAPP) performed well for small recombination rates. StarBEAST2 inferred more accurate divergence time at deeper timescale whereas SNAPP performed better estimation of divergence time at shallow divergence (Supplementary Fig. S1 and Fig. 2). All methods yielded more accurate ancestral population sizes for the shallow simulations where the most accurate method was SNAPP; in contrast, SNAPP tended to overestimate the population sizes backward in time for the deep simulations, reflecting the discussion of Bryant et al. (2012b) that "θ values can only be reliably inferred for ancestral populations if sufficiently many coalescent events occur within these populations."

It has been previously shown that recombination results in biased estimates of ancestral population parameters (Wall, 2003; Lohse and Frantz, 2014; Zhu et al., 2022). Here we repeated the StarBEAST2 analyses on data with growing segment lengths (200, 1000, and 5000 sites), as longer segments tend to encompass more intra-genic recombination breakpoints (Table 1). For small recombination rates, more sites slightly improved the estimates of StarBEAST2 (Supplementary Fig. S1 and Fig. 2). As recombination rate increases, the parameter estimates of longer segments started to deviate from the true values, with ancestral population sizes being underestimated but divergence times being overestimated. Remarkably, using sequences with 5000 sites under high recombination rates, StarBEAST2 underestimated the population sizes in terms of posterior mean by 9.5 % and 39 % on average, for the shallow and deep species trees, respectively. We noticed that such underestimation of population sizes was accumulated backwards in time (as reflected by $truth > N_{AB} > N_{ABC} > N_{ABCD}$). Despite a few exceptions (e.g., shallow phylogeny, $r = 2 \times 10^{-7}$, $\mu = 10^{-7}$, and length of 1000 sites), StarBEAST2 using short and intermediate segments seemed to be relatively robust to the presence of recombination. This supports the previous finding that excessive recombination events affected Bayesian analysis of multilocus sequence data assuming the MSC model (Zhu et al., 2022).

In comparison to StarBEAST2, SNAPP tended to overestimate N_{ABC} and N_{ABCD} at deep timescales. The performance of SNAPP, as expected, was in general unaffected by the recombination rate changes. As opposed to StarBEAST2, SNAPP yielded less variable ancestral population sizes at deep phylogenetic relationships.

Although diCal2 was designed to explicitly account for the recombination process, it appears to be the worst-performing one among all the methods evaluated here in this study. We found that diCal2 underestimated the root population size by about 43.2 % on average. Notably, we observed a large variance of diCal2's estimates between replicates for the deep phylogeny when the relative rates of recombination to mutation are 0.2 (N_{ABC}) and 2 (T_{AB} and T_{ABC}). To determine whether the observed large variance was due to convergence issues, we additionally conducted diCal2 analyses on the full data sets using narrow bounds for the population parameters. We only detected minor convergence problems under a model condition with high recombination rate and mutation rate ($r = 2 \times 10^{-7}$, $\mu = 10^{-7}$, and deep phylogeny; Supplementary Fig. S2). Hence, the observed erroneous behavior of diCal2 was not directly tied to convergence issues.

Table 1
Means and standard deviations of the average c-gene lengths of the simulated data. Each value was obtained from 10 replicates.

Recombination rate	Shallow		Deep	
	Mean	SD	Mean	SD
2×10^{-9}	1160.69	21.44	723.57	12.68
2×10^{-8}	117.37	0.98	73.00	0.32
2×10^{-7}	12.16	0.03	7.78	0.01

4. Discussion

The results presented here show that realistic levels of recombination are likely to have very little impact on Bayesian inference of evolutionary parameters under MSC using either multilocus data or unlinked SNPs. Surprisingly, anomalous behavior of diCal2, which was designed for analyzing full genome sequences in the presence of recombination, was found in our simulation study, most likely caused by extensive approximations employed in its algorithm. Still, the utility of MSC methods will depend on factors including the level of recombination, the lengths of loci, and the time scale of the phylogeny. The worse performance of StarBEAST2 when loci had thousands of sites clearly suggests that the amassing of intra-locus recombination events substantially hinders the power of MSC-based multilocus analysis. If the genetic loci were sampled from genomic regions with high recombination rates, it would be better to utilize larger numbers of short loci rather than fewer long loci. Our results also reinforce that the accuracy of SNAPP is impacted by the level of ILS: in the presence of extensive ILS, SNAPP yields highly accurate results, whereas in the absence or presences of low levels of ILS, SNAPP constitutes a biased estimator of population parameters.

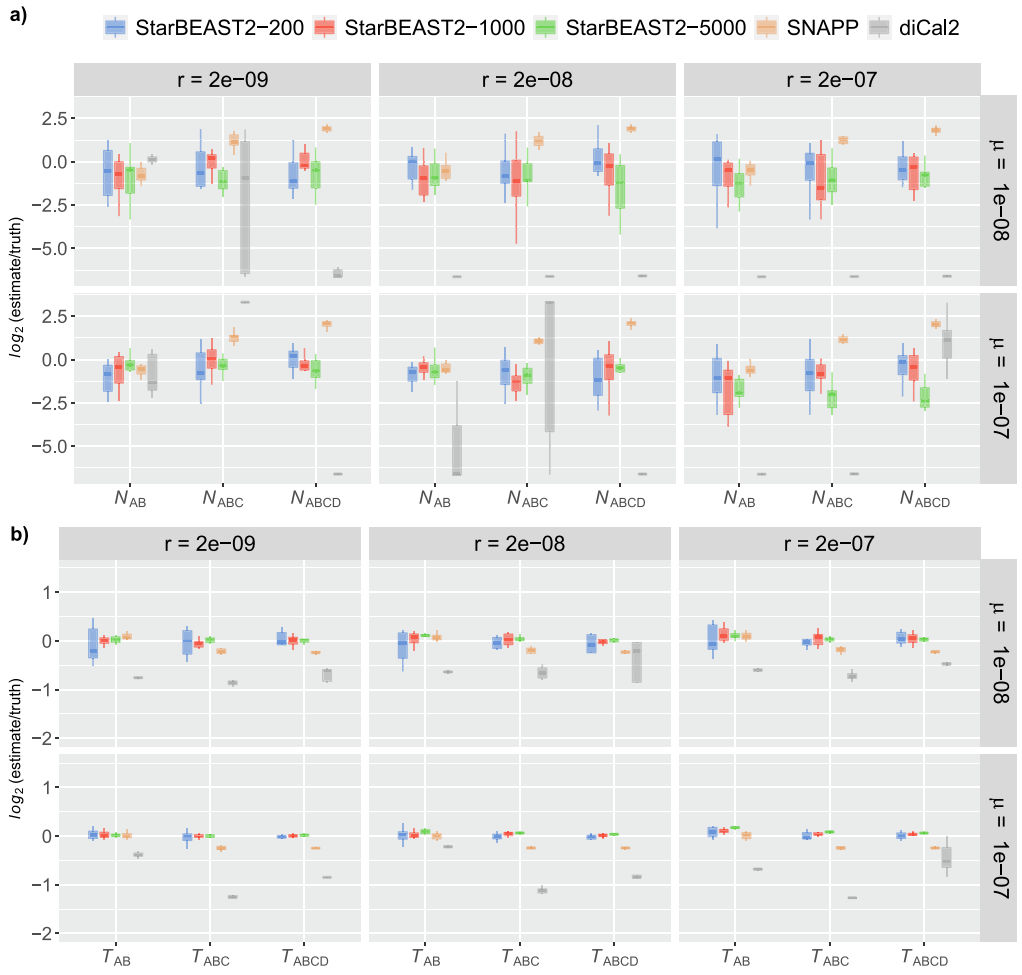

Fig. 2. Inference results on the deep phylogeny. Results are shown for five methods: StarBEAST2 with gene alignments of length 200 (StarBEAST2-200), 1000 (StarBEAST2-1000), and 5000 (StarBEAST2-5000), SNAPP, and diCal2. a) Boxplots showing the estimation error of population sizes. N_{AB} = the population size of the A-B ancestral population, N_{ABC} = the population size of the A-B-C ancestral population, and N_{ABCD} = the population size of the root population. b) Boxplots showing the estimation error of divergence times. T_{AB} = the time of the A-B split, T_{ABC} = the time of the A-B-C split, and T_{ABCD} = the time of the A-B-C-D split.

In terms of the running time of methods, due to differences in the underlying algorithms and variation in the chain lengths, the efficiency of the MCMC-based methods, namely StarBEAST2 and SNAPP, was measured by wall-clock time required per unit ESS so that the number of iterations is not a factor, whereas the efficiency of diCal2 was assessed by wall-clock time spent on a single data set. Using these measures, StarBEAST2 was faster than SNAPP in our study, taking 0.27 min on average to obtain 1 ESS compared to 0.51 min in the case of SNAPP. diCal2 was the least efficient method with an average runtime of 130 wall-clock hours on each data set. It is worth mentioning that diCal2 could be conducted in a parametric bootstrapping manner. But given the computational cost of diCal2, performing a bootstrap analysis would be prohibitive on our data sets.

While diCal2 failed to achieve desirable performance in our study, it is still worth exploring other inference techniques under the multispecies coalescent with recombination that utilize different mathematical or algorithmic frameworks. Furthermore, beyond the estimation of evolutionary parameters, StarBEAST2 has the advantage of being able to produce better estimates of gene trees than those obtained by methods that infer gene trees independently of the species phylogeny.

Our analyses of how recombination affects the estimation of ancestral population parameters is similar to the recent work by Zhu et al. (2022). However, the latter focused on the effect of recombination on Bayesian analyses for addressing different phylogenetic questions under the MSC. In contrast, our work explores the power of different types of methods under various evolutionary scenarios. Zhu et al. (2022) simulated data with intra-locus recombination under the multispecies network coalescent (Yu et al., 2014) model and varied the number of loci as well as number of sequences. We simulated whole-genome data under the MSC, and then sampled distantly-spaced segments of small, medium, and large sizes across the genome, which is a common practice in empirical phylogenomic analyses. As a result, our data included the effect of both intra- and inter-locus recombination, whereas Zhu et al. (2022)'s data incorporated intra-locus recombination and introgression, but not inter-locus recombination.

It is important to note that neither our study nor that of Zhu et al. (2022) incorporated variation of substitution or recombination rates among lineages and sites. Rate heterogeneity has been widely observed in real data (Nabholz et al., 2008; Beeson et al., 2019). In addition, it has been demonstrated that ILS could result in apparent substitution rate variation, which in turn may cause technical bias in evolutionary parameter estimation (Mendes and Hahn, 2016), and the use of species tree relaxed clocks enables more robust parameter inference (Ogilvie et al., 2017). Therefore, it is worth investigating how reliable current methods with or without relaxed clocks are in the presence of varying levels of rate heterogeneity.

Finally, while we limited our study to the impact of recombination on MSC-based species tree inference methods, other evolutionary processes such as gene flow and gene duplication and loss could have an even much larger impact on the performance of those inference methods. In particular, it is worth exploring how the interaction among these processes impact phylogenomic inference under the MSC.

CRediT authorship contribution statement

Zhi Yan: Investigation, Formal analysis, Data curation, Visualization, Writing – original draft. **Huw A. Ogilvie:** Supervision, Visualization, Writing – review & editing. **Luay Nakhleh:** Conceptualization, Supervision, Resources, Funding acquisition, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by NSF grants CCF-1514177, CCF-1800723 and DBI-2030604 to L.N.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2023.107724.

References

Beeson, S.K., Mickelson, J.R., McCue, M.E., 2019. Exploration of fine-scale recombination rate variation in the domestic horse. GenomeRes. 29, 1744–1752.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., RoyChoudhury, A., 2012a. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29, 1917–1932.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N.A., RoyChoudhury, A., 2012b. Inferring Species Trees Di- rectly from Biallelic Genetic Markers: Bypassing Gene Trees in a Full Coalescent Analysis. Mol. Biol. Evol. 29, 1917–1932. https://doi.org/10.1093/molbev/mss086.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., Mc- Carthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10. https://doi.org/10.1093/gigascience/giab008.

Doyle, J.J., 1997. Trees within Trees: Genes and Species, Molecules and Morphology. Systematic Biol- ogy 46, 537–553. https://doi.org/10.1093/sysbio/46.3.537.

Doyle, J.J., 2021. Defining Coalescent Genes: Theory Meets Practice in Organelle Phylogenomics. System- atic Biology 71, 476–489. https://doi.org/10.1093/sysbio/syab053.

Drummond, A.J., Suchard, M.A., Xie, D., Rambaut, A., 2012. Bayesian phylogenetics with beauti and the beast 1.7. Mol. Biol. Evol. 29, 1969–1973.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 7, 457–472.

Heled, J., Drummond, A.J., 2009. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27, 570–580.

Jensen-Seaman, M.I., Furey, T.S., Payseur, B.A., Lu, Y., Roskin, K.M., Chen, C.F., Thomas, M.A., Haussler, D., Jacob, H.J., 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14, 528–538.

Kelleher, J., Etheridge, A.M., McVean, G., 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. 12, e1004842.

Lanier, H.C., Knowles, L.L., 2012. Is Recombination a Problem for Species-Tree Analyses? Syst. Biol. 61, 691–701. https://doi.org/10.1093/sysbio/syr128.

Li, H., Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. Nature 475, 493–496.

Liu, X., Ogilvie, H.A., Nakhleh, L., 2021. Variational inference using approximate likelihood under the coalescent with recombination. Genome Res. 31, 2107–2119.

Lohse, K., Frantz, L.A., 2014. Neandertal admixture in eurasia confirmed by maximum-likelihood analysis of three genomes. Genetics 196, 1241–1251.

Mendes, F.K., Hahn, M.W., 2016. Gene tree discordance causes apparent substitution rate variation. Syst. Biol. 65, 711–721.

Nabholz, B., Glémin, S., Galtier, N., 2008. Strong variations of mitochondrial mutation rate across mammals—the longevity hypothesis. Mol. Biol. Evol. 25, 120–130.

Ogilvie, H.A., Bouckaert, R.R., Drummond, A.J., 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Mol. Biol. Evol. 34, 2101–2114.

Patel, S., Kimball, R.T., Braun, E.L., 2013. Error in phylogenetic estimation for bushes in the tree of life. J. Phylogenet. Evol. Biol 1, 1–10.

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.

Rieger, R., Michaelis, A., Green, M.M., 2012. Glossary of genetics and cytogenetics: classical and molecular. Springer Science & Business Media. Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L.B., Hood, L., Galas, D.J., 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science 328, 636–639. URL: https://www.science.org/doi/abs/10.1126/science.1186802, doi:10.1126/science.1186802, arXiv:https://www.science.org/doi/pdf/10.1126/science.1186802.

Roach, J.C., Glusman, G., Smit, A.F.A., Huff, C.D., Hubley, R., Shannon, P.T., Rowen, L., Pant, K.P., Goodman, N., Bamshad, M., Shendure,J., Drmanac, R., Jorde, L.B., Hood, L., Galas, D.J., 2010. Analysis of genetic inheritance in a family quartet by whole-genome sequencing.Science 328, 636–639. URL: https://www.science.org/doi/abs/10.1126/science.1186802, doi:10.1126/science.1186802,arXiv:https://www.science.org/doi/pdf/10.1126/science.1186802.

Springer, M.S., Gatesy, J., 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94, 1–33. https://doi.org/10.1016/j.ympev.2015.07.018.

Stange, M., Sánchez-Villagra, M.R., Salzburger, W., Matschiner, M., 2018. Bayesian divergence-time estimation with genome-wide snp data of sea catfishes (ariidae) supports miocene closure of the panamanian isthmus. Syst. Biol. 67, 681–699.

Steinrücken, M., Kamm, J., Spence, J.P., Song, Y.S., 2019. Inference of complex population histories using whole-genome sequences from multiple populations. Proceedings of the National Academy of Sciences 116, 17115–17120. URL: https://www.pnas.org/doi/abs/10.1073/ pnas.1905060116, doi:10.1073/pnas.1905060116, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.1905060116.

Wall, J.D., 2003. Estimating ancestral population sizes and divergence times. Genetics 163, 395–404.

Wang, Z., Liu, K.J., 2016. A performance study of the impact of recombination on species tree analysis. BMC Genomics 17, 165–174.

Yu, Y., Dong, J., Liu, K.J., Nakhleh, L., 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences 111, 16448–16453.

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 n/a. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16433, doi:https: //doi.org/10.1111/mec.16433, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.16433.

Zimmermann, T., Mirarab, S., Warnow, T., 2014. BBCA: Improving the scalability of *BEAST using random binning. BMC Genomics 15, 1–9.