Inference of Gene Flow from Genomic Data

Jiayi Ji

Supervisor: Prof. Ziheng Yang

Department of Genetics, Evolution and Environment

UCL

This dissertation is submitted for the degree of

Doctor of Philosophy

I, Jiayi Ji confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

July 23, 2025

Abstract

Species and populations did not evolve independently after splitting from their ancestors, and they were found to exchange alleles when coming into contact. The process of gene flow has been documented in numerous species throughout the tree of life. The exponential growth of genomic data over the past two decades has driven a surge in studies aiming to quantify the extent of gene flow across different systems and to understand the role of gene flow during and after speciation. Most efforts have been put on employing heuristic or approximate approaches that rely on summaries of sequence data, in which the rich information for inferring species divergence and cross-species gene flow is not fully leveraged and largely lost. Recent advances in the multispecies coalescent (MSC) model have made it a powerful framework for the inference of species tree and the estimation of two idealized formulations of gene flow: episodic introgression or continuous migration. These methods based on the MSC framework can capture more features of gene flow, including the strength, direction and timing, while also allowing for the estimation of key demographic parameters of speciation times and population sizes. This thesis focuses on gene flow inference based on the full likelihood methods implemented in Bayesian program BPP. We analyse genomic data from three different species systems. First, we apply the introgression model in Chapter 2 and both the introgression and migration models in Chapter 3 to re-analyse two previously generated datasets for chipmunk species group *Tamias quadrivittatus* and a *Drosophila* clade, identifying gene flow between both sister and non-sister species that summary methods failed to detect. Next, we compile three massive genomic datasets for chimpanzees and bonobos in Chapter 4, each of $>$ 50,000 loci. Model-based likelihood methods identify consistent migration events, whereas earlier evidence is mostly conflicting and geographically implausible. Lastly, in Chapter 5, we evaluate the impact of read depth on the inference of gene flow using coalescent-based methods through simulation and assess the influence of phasing in analysis of data at different depths. The work in the thesis highlights the importance of using statistically adequate methods to reach reliable biological conclusions concerning cross-species gene flow. The findings in the empirical data analysis imply that introgression is pervasive and not merely an exception in species evolution.

Impact Statement

Gene flow has been extensively identified across taxa, facilitated by methodological advances and revolutionary sequencing technologies. Despite the availability of high-quality genome assemblies, typically generated with substantial financial and time investment, a large proportion of published and ongoing genomic studies still evaluate gene flow using methods based on summaries of sequence data. The widespread use of these heuristic approaches, and the relative underutilization of full-likelihood methods, stem primarily from insufficient evaluation in realistic scenarios, regarding their power to detect gene flow. As a result, discussions of method performance often remain confined to simulations or small-scale empirical datasets.

We fully acknowledge that full-likelihood methods based on the multispecies coalescent framework tend to involve computationally intensive likelihood calculation, making their application to genome-scale datasets a valid concern. However, in this thesis, we demonstrate across multiple biological systems that statistically advanced methods are essential for accurately inferring gene flow among species and populations, including the direction, magnitude, and timing of gene flow, where applicable. In the thesis, we mainly address the following questions revolving full-likelihood, coalescent-based methods:

1. Whether full-likelihood approaches consistently outperform summary methods in empirical data analysis, as suggested by simulation studies.

2. Whether these methods are computationally efficient enough to handle realistically sized genomic datasets.

3. How these methods can be effectively and explicitly applied to real genomic data in practice.

4. What factors influence inference under the MSC framework, and in what ways they affect the results.

We highlight the superior statistical reliability of full-likelihood methods by analysing empirical data from several distinct species systems, including chipmunks (rodents), flies (insects), and chimpanzees and bonobos (primates). We show that it is now computationally feasible to apply these

methods to large-scale datasets consisting of thousands, or even tens of thousands, of loci that were previously limited to faster, summary-based approaches. Our work contributes to building a standard, easy-to-follow workflow for gene flow inference using methods implemented in BPP, encompassing multi-locus data compilation, species tree inference, gene flow model construction, and parameter estimation.

We also demonstrate that summary methods can suffer from multiple limitations. For example, in Chapters 2 and 3, methods such as HYDE and QUIBL are shown to detect only specific types of gene flow predefined in their models. When these model assumptions are violated, the power to detect gene flow is significantly compromised, which may lead to incorrect conclusions about the evolutionary history of gene exchange, which is further compounded by the information loss inherent in summary statistics. Our findings may serve as a reminder to users of these methods, encouraging them to understand the limitations and interpret results carefully, while also indicating the need to further improve their statistical properties in future.

Research Paper Declaration Form I

Title of the manuscript:	Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipmunks
DOI link:	https://doi.org/10.1093/sysbio/syac077
Journal:	Systematic Biology
Published date:	March 2023
Author(s):	Jiayi Ji, Donavan J Jackson, Adam D Leaché, Ziheng Yang
Peer review:	Yes
Copyright retained by author(s):	Yes
Statement of contribution:	Analyses were performed by Jiayi Ji. The distribution map was created by Donavan J Jackson and Adam D Leaché. Manuscript writing was done by Ziheng Yang.
Reused in the thesis's chapter(s):	Chapter 2

Candidate, date:

Supervisor/Senior author, date:

Research Paper Declaration Form II

Title of the manuscript:	Inference of Cross-Species Gene Flow Using Genomic Data Depends on the Methods: Case Study of Gene Flow in *Drosophila*
DOI link:	https://doi.org/10.1093/sysbio/syaf019
Journal:	Systematic Biology
Published date:	March 2025
Author(s):	Jiayi Ji, Thomas Roberts, Tomáš Flouri, Ziheng Yang
Peer review:	Yes
Copyright retained by author(s):	Yes
Statement of contribution:	Analyses of empirical data were performed by Jiayi Ji. The computer simulations were conducted by Thomas Roberts. The manuscript was written by Ziheng Yang.
Reused in the thesis's chapter(s):	Chapter 3

Candidate, date:

Supervisor/Senior author, date:

Acknowledgements

Embarking on my PhD journey four years ago was no easy feat. Yet, looking back, distilling the experiences and lessons of these years proves even more daunting.

I wanted to express my utmost gratitude to the renowned FRS Professor Ziheng Yang, whose enlightening and visionary supervision has guided me throughout the four-year study at UCL. His pre-eminent expertise in Bayesian phylogenetics and evolutionary biology have been navigating me through the complex scientific questions. Many of the ideas and hypotheses in my work would have never been conceived without him. What is beyond biology but equally important I have acquired from him is problem-solving skills, which have enabled me to handle unexpected situations or difficult challenges, not only in academia.

I would also like to thank the current and former members of the Yang lab for their tremendous intellectual and emotional support. In particular, I am grateful to Tomáš Flouri (for developing BPP and offering technical support on coding and statistics), Paschalia Kapli (for inspiring and contributing to the read-depth project), Sandra Álvarez-Carretero (for helping with practical modules), Thomas Roberts (for his simulation work in the *Drosophila* project). My research has substantially benefited from their exceptional, comprehensive skillsets in phylogenetics, computer science and statistics. I was also deeply touched by the laughter and joy in every cheering moment we shared and will always cherish their encouragement and support during my hard times.

I appreciated the collaboration within the cohort of the Department of Genetics, Evolution and Environment at UCL, and I would like to thank Harrion Ostridge and Prof. Aida Andrés for generously sharing chimpanzee genomic data and their input on data processing. I would like to thank Prof. Garrett Hellenthal for being my upgrade examiner and his constructive suggestions on my work. Additionally, I would like to extend my thanks for the contribution of our external collaborators: Prof. Adam Leaché (University of Washington) who offered incredible support on chipmunk biology, Prof. Bruce Rannala (UC Davis) who advised on genome assembly for chimpanzees and bonobos and the subsequent phylogenomic analysis.

I am thankful for my partner Yexin (Monica) Zhang as well as many amazing friends of mine in

London, Cambridge and China.

Lastly and most importantly, I love to dedicate the thesis to my beloved parents.

Contents

List of Figures

List of Tables

Chapter 1

Inference of Interspecific Gene Flow Using Genomic Data

1.1 The Prevalence and Impacts of Gene Flow

Over the past few decades, gene flow has moved from relative obscurity to a well-recognised component of evolution. Once considered to be rare, exchange of genetic material has now been identified in a wide variety of genera across the tree of life, such as *Homo* humans (Green *et al.*, 2010; Kuhlwilm *et al.*, 2016a; Li *et al.*, 2024), *Pan* chimpanzees and bonobos (Brand *et al.*, 2022; de Manuel *et al.*, 2016), *Heliconius* butterflies (Edelman and Mallet, 2021; Thawornwattana *et al.*, 2023b) and *Anopheles* mosquitoes (Fontaine *et al.*, 2015).

Basically, gene flow refers to the process by which alleles transfer from one population to another, genetically connecting geographically structured populations (Slatkin, 1985, 1987). As pertains to gene flow, the term *migration* is sometimes used synonymously, particularly when gene movement results from the dispersal of individuals. It was first introduced in the population genetics model *island model of migration* to describe the movement of alleles between populations through dispersal and reproduction (Wright, 1931). The concept *introgressive hybridization* or simply *introgression* was put forward and further elaborated by botanists as the cross-species infiltration of germ plasm through repeated backcrossing of hybrids to the parental species (Anderson, 1953; Anderson and Hubricht, 1938). The significant role of introgression had motivated botanists to find evidence of gene flow in multiple plant groups for decades (Arriola and Ellstrand, 1996; Brunsfeld *et al.*, 1992; Ellstrand, 2014; Heiser, 1947). Nevertheless, due to the limited availability of genetic markers for testing gene flow at the time, it was not very clear how common gene flow actually was in natural populations.

People have long doubted the existence of inter-species animal hybrids in nature, which were considered rare and largely maladaptive, suffering from hybrid sterility and inviability due to mismatched genomes (Mallet, 2005; Mayr, 1942, 1963, 1970). Thanks to the ever-expanding amount of genomic data, coupled with advances in the toolbox of methods for studying gene flow, the discovery has recently taken a big step forward. Today, cross-species introgression and hybridization have been

extensively documented, occurring far more frequently than previously thought (Adavoudi and Pilot, 2021; Mallet, 2007). Within-species gene flow was also found to be even more rampant, with alleles exchanged even between geographically separate populations (Sexton *et al.*, 2024).

The recognition of gene flow's role in evolution has accordingly undergone profound changes. Introgression may have a dynamic role as a homogenizing force and a barrier to divergence (Anderson, 1953; Rieseberg, 1997; Soltis and Soltis, 2009) or a catalyst for speciation and ecological adaptation (Baack and Rieseberg, 2007; Marques *et al.*, 2019; Taylor and Larson, 2019), depending on the genomic landscape. Recent studies have revealed its evolutionary implications in driving species diversification and accelerating local adaptation (Feder *et al.*, 2012; Folk *et al.*, 2018; Martin and Jiggins, 2017). Gene flow essentially transforms the traditional bifurcating tree-like view of evolution, calling for a network-like resolution of evolutionary history where species boundaries are blurred (Mallet *et al.*, 2016). This shift necessitates empirical studies across a wide range of organisms to build deeper insights into these aspects.

In this thesis, I apply full-likelihood methods to identify the presence and quantify the extent of introgression/migration using genomic data in systems including *Tamias quadrivittatus* chipmunks (Chapter 2), *Drosophila* flies (Chapter 3), and chimpanzees and bonobos of *Pan* genus (Chapter 4), in comparison to the previously used methods based on data summaries. In addition, computer simulation is conducted to assess the impact of sequencing read depth on the inference of species tree and gene flow in Chapter 5.

1.2 Overview of Methods for Gene Flow Inference

The toolkit for gene flow detection comprises methods of varying complexity, ranging from simple tests like the D-STATISTIC to sophisticated model-based approaches for inferring phylogenetic networks.

The most used methods include those relying on summaries of multilocus sequence data, e.g., site pattern counts (Blischak *et al.*, 2018; Green *et al.*, 2010), estimated gene trees (Solis-Lemus and Ane, 2016; Wen *et al.*, 2016) and site frequency spectra (Excoffier *et al.*, 2013; Gutenkunst *et al.*, 2009), which are referred to as *summary methods*. These methods tend to be heuristic and

make approximations in likelihood calculations, making them computationally efficient approaches primarily used for testing the presence of gene flow or estimating its strength under a predefined gene flow model with a fixed data setup. They are fairly popular in phylogenomic studies but prone to significant issues of model identifiability resulting from information loss (Pang and Zhang, 2024; Xu and Yang, 2016).

In contrast, methods that construct and evaluate the complete joint likelihood function of all observed data under a user-specified model are referred to as *full-likelihood methods*. Unlike summary-based methods, which use reduced data representations, full-likelihood methods leverage all available information in the data and make inference based on a likelihood function that connects sequence data to the underlying evolutionary processes (e.g., speciation history and gene-flow events). Nevertheless, the methods are much more computationally demanding and usually not reckoned as the optimal choice for genome-scale datasets. Recent algorithmic optimizations and progress in computer hardware have made it feasible to analyse datasets of thousands of genomic segments. In the first three chapters, I study the history of gene flow in a few species groups by analysing their genomic datasets containing $\sim$ 1000 to 500,000 loci using the full-likelihood approaches implemented in Bayesian program BPP (Flouri *et al.*, 2020, 2023; Yang, 2015).

1.2.1 Full-Likelihood Methods

The coalescent process and the multispecies coalescent model The mathematical theory of coalescent, first developed by Kingman (1982), describes the variation in the genealogical history of a sample of DNA sequences from one population, assuming the absence of selection, recombination and gene flow. While traditional population genetics has mainly focused on changes in allele frequencies, the emergence of the coalescent model presents a more natural approach for reconstructing demographic histories through the analysis of sequence data.

The multispecies coalescent (MSC) model extends the single-population framework for multiple species, which integrates the phylogenetic process of species divergences and the population genetic process of coalescent (Rannala and Yang, 2003). Given a species tree, the MSC model gives how genealogies from different species vary across different genomic regions. When sequences sampled

at a locus are traced backwards in time, coalescent events are Poisson processes occurring at a rate of $\frac{1}{2N}$, inversely proportional to the effective population size N. Define one time unit as the time to accumulate one mutation per site, any two lineages should coalesce at rate $\frac{2}{\theta}$, where $\theta = 4N\mu$ is the effective population size measured in the expected number of mutations per site. When there are > 2 lineages in the population, two sequences are chosen uniformly at random to coalesce. The coalescent waiting times follow an exponential distribution with rate $\binom{n}{2}\frac{2}{\theta}$, where n is the number of lineages that have not yet coalesced at that time. It might be difficult for sequences to coalesce in time within large populations that existed for short periods in history, as reflected by short branch length in coalescent units $\frac{2\tau}{\theta} = \frac{T}{2N}$. Uncoalesced sequences should enter the next parental population after they reach the end of the current population as specified by species divergence time $\tau = T\mu$, measured in the number of mutations, and T is the absolute divergence time in generations. Hence, it is naturally accommodated in the MSC framework that sequences do not necessarily coalesce as soon as they reach the most recent common ancestor tree but instead coalesce in more ancient ancestors, which is known as *incomplete lineage sorting* or *deep coalescence*. The coalescent process ends if all sequences have coalesced into a single lineage.

The inference framework of the MSC model operates directly on sequence data. Let $D = \{D_i\}$ be the sequence data, where D_i represents the sequence alignment at ith locus for $i = 1, 2, 3, \ldots, L$. The species tree (S, Θ) is defined by the topology S with parameter vector Θ, including parameters of species divergence times τ, population sizes θ. Also, let $G = \{G_i\}$ and $t = \{t_i\}$ be the gene trees and coalescent times (or branch lengths), where the gene tree at the locus i is G_i with coalescent times t_i. We have the likelihood of the sequence data and the gene tree at any single locus i

$$f(D_i, G_i, t_i \mid S, \Theta) = f(D_i \mid G_i, t_i) f(G_i, t_i \mid S, \Theta), \tag{1.1}$$

where the phylogenetic likelihood $f(D_i \mid G_i, t_i)$ given the gene tree (G_i, t_i) can be calculated assuming a time-reversible substitution model with Felsenstein's pruning algorithm (Felsenstein, 1981), and the density $f(G_i, t_i \mid S, \Theta)$ is specified by the multispecies coalescent process (Rannala and Yang, 2003). Note that the data D_i and the species tree (S, Θ) are statistically independent given the gene tree (G_i, t_i). However, the gene tree (G_i, t_i) at each locus is not observed, so the likelihood should sum over all

possible gene tree topologies and integration over the coalescent times at each locus:

$$f(D_i \mid S,\Theta) = \sum_{G_i} \int_{t_i} f(D_i \mid G_i,t_i) f(G_i,t_i \mid S,\Theta)\, dt_i. \tag{1.2}$$

Assuming gene trees at different loci are independent, the likelihood of data $D = \{D_i\}$ is simply a product of likelihoods across loci:

$$f(D \mid S,\Theta) = \prod_{i=1}^{L} f(D_i \mid S,\Theta). \tag{1.3}$$

Equation 1.3 is the inference basis for maximum likelihood (ML) methods. In contrast, Bayesian inference is typically performed based on the joint posterior distribution, using a Markov chain Monte Carlo (MCMC) algorithm to average over gene trees:

$$\begin{aligned} f(S,\Theta,G,t \mid D) &\propto \pi(S,\Theta) f(D,G,t \mid S,\Theta) \\ &= \pi(S,\Theta) \prod_{i=1}^{L} f(D_i \mid G_i,t_i) f(G_i,t_i \mid S,\Theta), \end{aligned} \tag{1.4}$$

where $\pi(S,\Theta)$ is the prior on species tree topology and demographic parameters. The MSC density $f(G_i,t_i \mid S,\Theta)$ in eq. 1.4 is straightforward to be derived by traversing the populations and examining the coalescent events. For example, the gene tree given the MSC model in figure 1.1a has probability

$$\begin{aligned} f(G_i,t_i \mid S,\Theta) = &\frac{2}{\theta_A} e^{-\frac{2}{\theta_A} t_1} && \text{(Population A)} \\ &\times e^{-\frac{2}{\theta_B} \tau_T} && \text{(Population B)} \\ &\times e^{-\frac{2}{\theta_C} \tau_T} && \text{(Population C)} \\ &\times \frac{2}{\theta_T} e^{-\frac{12}{\theta_T}(t_2-\tau_T)} \times \frac{2}{\theta_T} e^{-\frac{6}{\theta_T}(t_3-t_2)} \times e^{-\frac{2}{\theta_T}(\tau_R-t_3)} && \text{(Population T)} \\ &\times \frac{2}{\theta_R} e^{-\frac{6}{\theta_R}(t_4-\tau_R)} \times \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_5-t_4)}. && \text{(Population R)} \end{aligned} \tag{1.5}$$

For species A, the contribution is $\frac{2}{\theta_A} e^{-\frac{2}{\theta_A} t_1}$, as there was one coalescent event between a_1 and a_2 at time t_1. In species B, there was no coalescent, so the probability of having no coalescent when there were two sequences during the time periods $(0,\tau_T)$ is $e^{-\frac{2}{\theta_B}\tau_T}$. At time t_2, it randomly chose two from four lineages in species T to coalesce so the rate is $\binom{4}{2}\frac{2}{\theta_T} = \frac{12}{\theta_T}$, and so forth at time t_3.

Although both ML and Bayesian frameworks enable joint estimation of the species phylogeny S and parameters in Θ or parameters solely on a fixed species tree, Bayesian inference (eq. 1.4) may be more efficient than ML methods (eq. 1.3) in practice since the multi-dimensional integral in eq. 1.2 is almost impossible to calculate except for small datasets. Bayesian statistics assign prior distributions to parameters in the model. With a conjugate prior used for a given parameter, the posterior is tractable and that parameter is allowed to be integrated out. For example, in species tree inference, population sizes θs may be integrated out analytically through the use of inverse-gamma priors, which improves MCMC mixing (Hey and Nielsen, 2007).

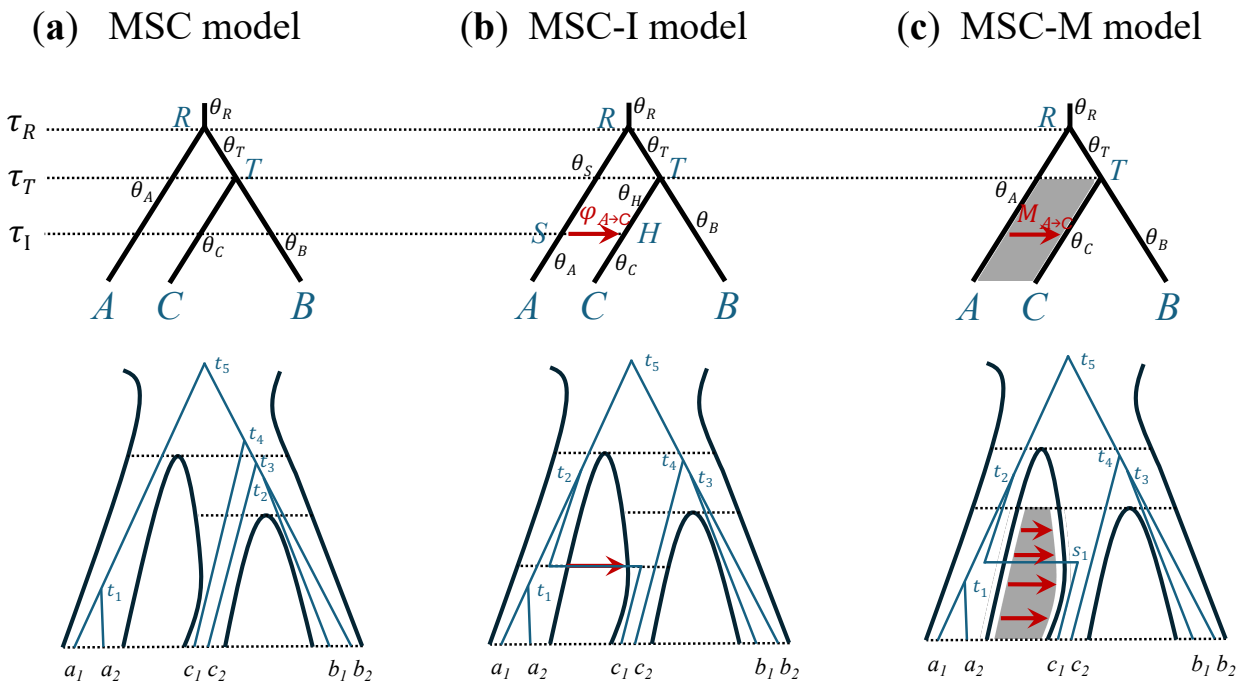

Figure 1.1: The multispecies coalescent (MSC) model and its gene flow extensions, the MSC-I and MSC-M models. The digrams in the top row include an instance of species tree (S,Θ) under the MSC, MSC-I and MSC-M models. The diagrams below display a possible coalescent process (G,t) of six sequences (a_1, a_2 from A; b_1, b_2 from B; and c_1, c_2 from C) given the species tree (S,Θ). **(a)** MSC model with no gene flow. Sequences are within the species tree, and those from different species can coalesce in the common ancestors. **(b)** MSC-I model with introgression from A to C (in forward time) at time τ_I with introgression probability $\varphi_{C\to B}$. **(c)** MSC-M model with continuous gene flow from A to C (in forward time) at a constant rate of $M_{A\to C}$ throughout the time period $(0,\tau_T)$. From a backward-in-time perspective, sequence c_1 was migrated from C to A at time s_1. The tip side indicates present time.

More recently, the MSC model has been generalized to incorporate gene flow, accommodating it in either episodic introgression or continuous-time migration, which opens the door for model-based inference of gene flow using genomic data. Currently, there are multiple Bayesian implementations

of full-likelihood methods aware of gene flow, such as BPP MSC-I (Flouri *et al.*, 2020) and MSC-M (Flouri *et al.*, 2023), *BEAST (Heled and Drummond, 2010), PHYLONET MCMC_SEQ (Wen and Nakhleh, 2018), G-PHOCS (Gronau *et al.*, 2011) and IMA3 (Hey *et al.*, 2018).

The multispecies coalescent with introgression (MSC-I) The first type of gene-flow model, the *multispecies coalescent model with introgression* (MSC-I), specifies pulses of gene flow (Flouri *et al.*, 2020; Yu *et al.*, 2014). The model assumes episodic gene flow in the past, with hybridization taking place between populations or species within a short period of contact. The MSC-I model is also known as the *multispecies network coalescent* (Wen *et al.*, 2016; Yu *et al.*, 2012) or *network multispecies coalescent* (Degnan, 2018) for its setting on species networks.

On top of having nodes representing speciation, the MSC-I model introduces hybridization nodes on species trees. Each hybridization node has two parents, with one representing the backbone of the species tree, and the other connected through introgression. When one sequence is traced backwards in time, it may encounter hybridization events and decide on which way to go, and it is possible for the sequence to transfer into another population at the time. Each introgression event is defined by two parameters: introgression time τ_I and introgression probability φ. The introgression probability $\varphi_{X\to Y}$ is the proportion of lineages in species Y that come from species X at the introgression time, forward in time. If traced backwards, for example, in the MSC-I model of figure 1.1b, species A received one of the lineages from C through the introgression A to C at time τ_I. The term *introgression probability* used in BPP is synonymous with the *inheritance probability* of Yu *et al.* (2014) and the *heritability* of Solis-Lemus and Ane (2016).

The MSC-I model is available in BPP (Flouri *et al.*, 2020), BEAST2 SPECIESNETWORK (Zhang *et al.*, 2018), and PHYLONET (Wen and Nakhleh, 2018). The Bayesian implementation in BPP includes four variants of MSC-I models that can be used for different introgression scenarios: unidirectional (fig. 1.2b) and bidirectional (fig. 1.2d) gene flow, and hybrid speciation (fig. 1.2c), in which the donor populations are still alive today. It is also possible that one of the parental populations or both went extinct after hybridization, as shown in figure 1.2a. Essentially, BPP does not implement MCMC moves that change the MSC-I model but estimate parameters under a fixed model. There are

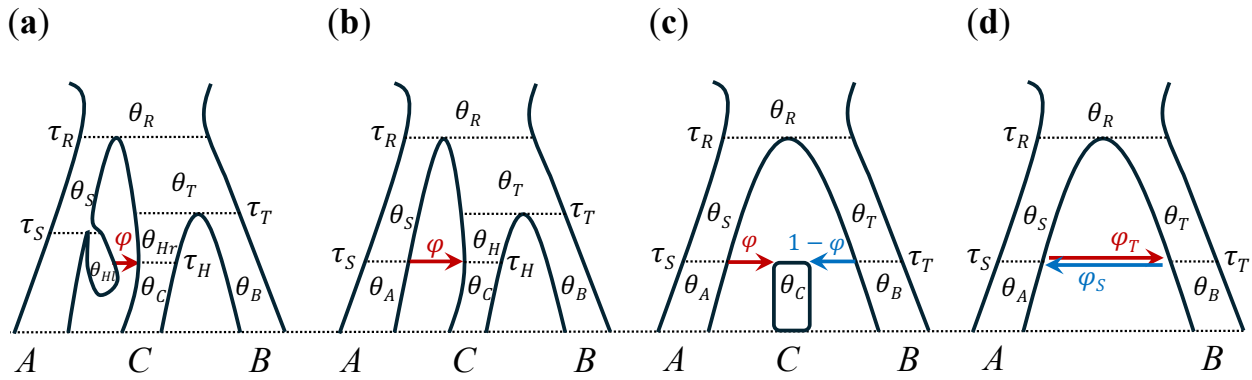

Figure 1.2: Four introgression models in BPP. **(a)** Introgression from an extinct or unsampled parental species H_l. **(b)** Unidirectional introgression from species A to C. **(c)** Hybrid speciation model. Two parental species A and B contacted with each other and gave birth to hybrid species C. **(d)** Bidirectional introgression between sister species A and B. Modified from figure 1 in Flouri *et al.* (2020).

few methods that can infer introgression models by allowing changes to hybridization events in the MCMC, such as PHYLONET MCMC_SEQ (Wen and Nakhleh, 2018). However, the MCMC does not mix efficiently, and they are barely applied to realistically sized datasets.

The calculation of density $f(G_i, t_i \mid S, \Theta)$, similarly, can be achieved by examining the occurrence of coalescent in each population. As for introgression, each time when a sequence passes a hybridization node, there is a probability φ or $1-\varphi$ depending on which route it has taken. For the gene tree of figure 1.1b, we have

$$\begin{aligned} f(G_i, t_i \mid S, \Theta) = & \frac{2}{\theta_A} e^{-\frac{2}{\theta_A} t_1} && \text{(Population A)} \\ & \times e^{-\frac{2}{\theta_B} \tau_T} && \text{(Population B)} \\ & \times e^{-\frac{2}{\theta_C} \tau_I} && \text{(Population C)} \\ & \times \varphi_{A\to C} \times \frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_2-\tau_I)} && \text{(Population S)} \\ & \times (1-\varphi_{A\to C}) && \text{(Population H)} \\ & \times \frac{2}{\theta_T} e^{-\frac{6}{\theta_T}(t_3-\tau_T)} \times \frac{2}{\theta_T} e^{-\frac{2}{\theta_T}(t_4-t_3)} && \text{(Population T)} \\ & \times \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_5-\tau_R)}. && \text{(Population R)} \end{aligned} \tag{1.6}$$

Sequence c_1 took the introgression path at time τ_I, which then coalesced in species S at time t_2, so that the contribution to the gene tree density from S is $\varphi_{A\to C} \times \frac{2}{\theta_S} e^{-\frac{2}{\theta_S}(t_2-\tau_I)}$. The other sequence c_2 took

the parental path and stayed in species C. Since that there was only one lineage in species H (after time τ_I and before τ_T, the contribution of the species is $1-\varphi_{A\to C}$. The terms of coalescent events can be derived using the same way as in the MSC model.

The multispecies coalescent with migration (MSC-M) Gene flow may occur over a prolonged period in nature. The second type of model assumes continuous gene flow occurring at a constant rate every generation over an extended time period (Hey and Nielsen, 2004; Nielsen and Wakeley, 2001), and it is referred to as the *multispecies coalescent model with migration* (MSC-M) (Flouri *et al.*, 2023) or *isolation with migration* (IM) model (Chung and Hey, 2017; Hey, 2010b).

In the MSC-M model, migration is specified under the idealized assumption of constant gene flow until the end of coexistence of involved populations. For example, species A in figure 1.1c has been sending migrants to population C at rate $M_{A\to C}$ since their split at time τ_T. The population migration rate $M_{X\to Y} = N_Y m_{X\to Y}$ is defined as the number of individuals moved from population X to Y per generation with time running forwards, where $m_{X\to Y}$ is the proportion of migrants in Y from X and N_Y is the number of individuals in recipient population Y. Some variants of the model are flexible with the spanning time of migration and allow it to be specified within a certain time frame rather than throughout the entire contemporary periods (Costa and Wilkinson-Herbots, 2017).

Two strategies are developed to calculate the density of gene trees $f(G_i, t_i \mid S, \Theta)$ under the MSC-M model, which differ in whether the migration history is integrated out. The first strategy relies on the *structured coalescent* (Nath and Griffiths, 1993; Notohara, 1990; Takahata, 1988; Wilkinson-Herbots, 1998). If the gene tree at a locus is represented by the topology and the coalescent times without including the migration history, the coalescent process can be described using a continuous-time Markov chain, and states in the chain consist of all possible configurations of the number and location of sequences (including the coalesced ones) (Andersen *et al.*, 2014; Zhu and Yang, 2012). Consider the simplest case of two sequences a and b sampled from two species A and B, respectively. In the MSC-M model with migration in both directions, the state of the two sequences before they enter the root must be one of the following: (1) one in species A and the other in species B, (2) both in species A, (3) both in species B, or (4) a and b coalesced in species A or B. In either species

A or B, a pair of sequences coalesces at rate $\frac{2}{\theta}$, and each sequence can migrate backward in time to another population at rate $\frac{4M}{\theta}$. The density calculation can then be achieved by constructing a transition matrix for each of time periods, in which the rates of coalescent and migration are constant. In practice, it is efficient only for a small number of species and sequences, as the state space expands exponentially as the number of species/sequences increases (Hobolth *et al.*, 2011). For example, the density $f(G_i, t_i \mid S, \Theta)$ in figure 1.1c is given as:

$$\begin{aligned} f(G_i, t_i \mid S, \Theta) =& P(t_1)_{AACC,ACC} \times P(\tau_T - t_1)_{ACC,AAC} && \text{(Population A and C)} \\ & \times \frac{2}{\theta_A} e^{-\frac{2}{\theta_A}(t_2 - \tau_R)} && \text{(Population A)} \\ & \times e^{-\frac{2}{\theta_B}\tau_T} && \text{(Population B)} \\ & \times \frac{2}{\theta_T} e^{-\frac{6}{\theta_T}(t_3 - \tau_T)} \times \frac{2}{\theta_T} e^{-\frac{2}{\theta_T}(t_4 - t_3)} && \text{(Population T)} \\ & \times \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_5 - \tau_R)}. && \text{(Population R)} \end{aligned} \tag{1.7}$$

The term $P(t_1)_{AACC,ACC}$ is the transition probability over $(0, t_1)$, from the initial state that two sequences in species A and two in C to the state where the sequences in A have coalesced while the sequences in C remain still. Similarly, $P(\tau_T - t_1)_{ACC,AAC}$ represents the transition over (t_1, τ_T) where one sequence from species C migrates into species A, but the migrated sequence has not coalesced with the one in A. The formulation is implemented in ML method 3s that infers migration rates using 3 sequences from 3 species (Dalquen *et al.*, 2017; Zhu and Yang, 2012). In Chapter 4, the method is applied to evaluate the gene flow among chimpanzee subspecies and bonobos.

Alternatively, the gene tree at each locus may include a full history of both coalescent and migration, which is implemented in G-PHOCS (Gronau *et al.*, 2011), IMA3 (Hey *et al.*, 2018), BEAST2 DENIM (Jones, 2019) and BPP MSC-M (Flouri *et al.*, 2023). Both coalescent events and migration events are described using Poisson distributions, with rate $\frac{2}{\theta}$ and $\frac{4M}{\theta}$ (Beerli and Felsenstein, 2001; Wang and Hey, 2010). Although including migration history in gene trees may slightly increase computational cost, the derivation of probability density with this strategy is much more straightforward, applicable for an arbitrary number of species and samples and suitable for Bayesian inference in most cases. To calculate the density, we can divide the entire timeline into multiple segments, each

with fixed populations and migration events, so that the rates of these events remain constant within each period (Jiao *et al.*, 2021). For the instance of the MSC-M model in figure 1.1c, the density $f(G_i, t_i \mid S, \Theta)$ is derived as:

$$\begin{aligned} f(G_i, t_i \mid S, \Theta) = & \frac{2}{\theta_A} e^{-\frac{2}{\theta_A} t_1} \times \frac{2}{\theta_A} e^{-\frac{2}{\theta_A}(t_2 - s_1)} && \text{(Population A)} \\ & \times e^{-\frac{2}{\theta_B} \tau_T} && \text{(Population B)} \\ & \times e^{-\frac{2}{\theta_C} s_1} \times \frac{4M_{A\to C}}{\theta_C} e^{-\frac{4M_{A\to C}}{\theta_C}(2s_1 + \tau_T - s_1)} && \text{(Population C)} \\ & \times \frac{2}{\theta_T} e^{-\frac{6}{\theta_T}(t_3 - \tau_T)} \times \frac{2}{\theta_T} e^{-\frac{2}{\theta_T}(t_4 - t_3)} && \text{(Population T)} \\ & \times \frac{2}{\theta_R} e^{-\frac{2}{\theta_R}(t_5 - \tau_R)}. && \text{(Population R)} \end{aligned} \tag{1.8}$$

There is a migration event A to C during the time interval $(0, \tau_T)$. When traced backwards, one lineage in species C moved into species A at time s_1 and the other one remained in C throughout the period. Their migration components in the density are given by $\frac{4M_{A\to C}}{\theta_C} e^{-\frac{4M_{A\to C}}{\theta_C} s_1}$ and $e^{-\frac{4M_{A\to C}}{\theta_C} \tau_T}$, respectively. The terms of coalescent events can be derived same as above.

In practice, it may be biologically sensible to estimate migration rates $M = Nm$ in the MSC-M model as a measure of migration intensity. Statistically speaking, expressing migration rate parameters as $W = \frac{4M}{\theta}$ makes it possible to sample directly from the posterior with conjugate priors assigned, which enables Gibbs sampling and improves the mixing of the MCMC algorithm. Methods such as BPP and IMA3 implement the parametrization to facilitate MCMC. For migration rates $W = \frac{4M}{\theta} = \frac{m}{\mu}$ measured on a mutation scale, one time unit is the time to accumulate one mutation per site, consistent with that used for divergence times τ and population sizes θ.

1.2.2 Summary or Heuristic Methods

In addition to likelihood methods, there are a variety of heuristic methods developed to study gene flow. Various data summaries, such as site-pattern counts, estimated gene trees and site frequency spectra (SFS), are efficiently reduced representation of multi-locus sequence data, and the relevant inference frameworks can be viewed as approximation of the full-likelihood methods detailed above. The simplicity can be a double-edged sword. It simplifies the statistical framework and reduces the

computational cost by bypassing the expensive likelihood calculations, but it fails to exploit all data information, so data summaries are typically not sufficient statistics for resolving all model parameters (Jiao *et al.*, 2021; Xu and Yang, 2016). For example, one of the well-known limitations among most of summary methods is that they have no power to identify gene flow between sister species, because of the lack of information in the variation of genealogical histories. Furthermore, for gene flow between non-sister species, they may have low statistical power to detect its presence, and some methods do not provide an estimator of strength and direction (Ji *et al.*, 2023). The power comparison, full-likelihood versus summary methods, is one of the key topics covered in the thesis. These methods have been extensively used and evaluated (Hibbins and Hahn, 2022; Ji *et al.*, 2023; Pang and Zhang, 2024). There is no denying that they have a role in preliminary evaluation of introgression signals, especially on phylogenies of many species. However, by themselves, it is rather difficult to accurately reconstruct the detailed history of gene flow. Here is a classification of major summary methods according to the type of inputs.

Counts of parsimony-informative site patterns and SNP data Genome-wide scans of population genetic statistics have become a common strategy for detecting signatures of gene flow between populations and species. Parsimony site-pattern counts are a typical representative of these statistics. Consider a quartet tree $(((S_1,S_2),S_3),O)$ of three species S_1, S_2, S_3 and outgroup O. Suppose one individual (or one lineage) sampled per species, for a single nucleotide site, biallelic site pattern *BBAA* matches the branching order of the species tree, where the ancestral allele A is possessed by individuals from species S_3 and O, and the derived allele B is possessed by species S_1 and S_2. Under the null hypothesis of no gene flow, the frequencies or counts of mismatching patterns *ABBA* and *BABA* pooled across genome are expected to be equal. Deviations from this null expectation are interpreted as evidence for gene flow between non-sister species S_1 and S_3 or S_2 and S_3. This is the widely applied D-STATISTIC (Durand *et al.*, 2011a; Green *et al.*, 2010). There are multiple variants of the statistic, such as D_{foil} (Pease and Hahn, 2015), which extends the framework to five taxa and allows for the detection of introgression directionality, and HYDE (Blischak *et al.*, 2018; Kubatko and Chifman, 2019), which can be used to estimate introgression probabilities under a predefined hybrid-speciation

model. These methods, however, are documented to be sensitive to low effective population sizes and window sizes (Martin *et al.*, 2015; Zheng and Janke, 2018). Almost all these methods implicitly assume equal population sizes and the infinite-sites model of mutations, frequently challenged in real data analysis.

The related F-STATISTICS (Patterson *et al.*, 2012) have also been actively applied for studying population admixture in a range of contexts. The F-STATISTICS are mainly based on SNP data and involve $f_2(A,B)$, $f_3(A;B,C)$ and $f_4(A,B;C,D)$ statistics that measure shared genetic drift among 2, 3 or 4 populations, respectively. The $f_4(A,B;C,D)$ statistic quantifies the covariance in allele frequencies between two population pairs (A,B) and (C,D) in an unrooted tree, and it is the same as the D-STATISTIC up to a normalization factor. The f_4-ratio statistic (Patterson *et al.*, 2012) can estimate the genome-wide mixing proportion in an admixed species or population, with similar functionality as HYDE (Blischak *et al.*, 2018). The f_d (Martin *et al.*, 2015) and its close relative f_dM (Malinsky *et al.*, 2015) are modified versions of the f_4-ratio, designed to detect local introgression signals in short genomic regions. The f-branch (or f_b) statistic (Malinsky *et al.*, 2018) attempts to assign introgression events to specific internal branches on a phylogeny using f statistics calculated based on species triplets. Most of the statistics mentioned above are integrated in the toolbox DSUITE (Malinsky *et al.*, 2021), which provides efficient computation of D-statistics and f_4-ratios across trios. It is also possible to infer the admixture history from the summary statistics. For example, the graph-based method QPGRAPH (Maier *et al.*, 2023; Patterson *et al.*, 2012) aims to find the optimal admixture graph that minimizes the error between fitted and estimated f statistics. Methods such as TREEMIX (Pickrell and Pritchard, 2012) and ORIENTAGRAPH (Molloy *et al.*, 2021) conduct inference under similar ideas.

One fundamental issue for the class of methods is that when site pattern counts are averaged across the genome, gene tree branch lengths at different loci become indistinguishable (Lohse and Frantz, 2014; Zhu and Yang, 2021). The variation in branch lengths provides crucial information about divergence times, and without it, it is impossible to detect gene flow between sister species (Ji *et al.*, 2023; Jiao *et al.*, 2021).

Estimated gene trees topologies The principle underlying gene flow tests based on gene tree topologies is the same as that behind methods using nucleotide site patterns, as both aim to detect inequalities in summary statistics to infer introgression between non-sister species. Given species tree $((S_1,S_2),S_3)$, introgression between S_1 and S_3 or S_2 and S_3 alters the expected equivalence in the probabilities of gene trees $((s_1,s_3),s_2)$ and $((s_2,s_3),s_1)$, causing one to occur at a higher frequency than the other. In the MSC-I model, the probabilities of three possible gene tree topologies can be analytically derived as functions of introgression probabilities (φ), population sizes (θ) and split times (τ) (Jiao *et al.*, 2021; Solis-Lemus and Ane, 2016). This is implemented in SNaQ (Solis-Lemus *et al.*, 2017) to estimate inheritance probabilities γ, conceptually synonymous to introgression probabilities φ in BPP. The toolkit PhyloNet includes two methods for estimating gene flow given a list of gene tree topologies — ML module InferNetwork_ML (Yu *et al.*, 2013, 2014) and its maximum pseudolikelihood version InferNetwork_MPL (Cao and Nakhleh, 2019; Yu and Nakhleh, 2015). Bayesian inference is also available in module MCMC_GT (Wen *et al.*, 2016). Besides, the discordant-count test (DCT) (Suvorov *et al.*, 2022) relies on a model of three species that implicitly assumes one-way introgression, but the introgression proportion estimate in DCT is biased. PHRAPL evaluates migration models by comparing the simulated tree topologies under a given model and the estimated topologies from data (Jackson *et al.*, 2017). Overall, the power of topology-based methods is essentially associated with the phylogenetic errors in gene tree reconstruction, which may be considerable for closely related species. They may struggle to resolve the history of gene flow in the presence of ghost introgression from unsampled lineages due to identifiability issues (Pang and Zhang, 2024).

Gene tree branch lengths Gene flow may leave signatures on branch lengths in gene trees. If there is gene flow between non-sister species, the branch lengths of gene trees $((s_1,s_3),s_2)$ and $((s_2,s_3),s_1)$, which support introgression, are expected to be shorter than those in tree $((s_1,s_2),s_3)$, as introgressed alleles tend to coalesce more recently than those following the species tree topology. QuIBL (Edelman *et al.*, 2019) assumes a 3-species inflow introgression model (with gene flow from the outgroup to one of the ingroups) and distinguishes introgression from incomplete lineage sort-

ing using gene tree branch lengths. There are several other methods devised similarly, such as the branch-length test (BLT) (Suvorov *et al.*, 2022), DIP (Forsythe *et al.*, 2020) and the statistics detailed in Hibbins and Hahn (2019). For instance, BLT infers introgression by statistically testing the differences in coalescent times on the two discordant gene trees, assuming unidirectional introgression in either way. Methods based on branch lengths may be inconsistent because estimated gene tree branch lengths are more susceptible to sampling error (DeGiorgio and Degnan, 2014).

Site frequency spectra The site frequency spectrum (SFS) summarizes the distribution of allele frequencies at all biallelic sites (or only biallelic SNPs) across a given set of regions within a single population or among few populations, which is also referred to as the allele frequency spectrum (AFS). It tends to sample multiple individuals per population. In the context of single population, the 1D-SFS contains information concerning the effective population size. As for the multi-species SFS, it informs interspecific parameters such as species divergence and introgression probabilities/migration rates. Wide debate has arisen regarding the information content of the SFS for inferring demographic histories, with discussions centred on whether distinct historical processes can produce identical allele frequency distributions (Baharian and Gravel, 2018; Bhaskar and Song, 2014; Myers *et al.*, 2008; Terhorst and Song, 2015).

SFS-based methods infer parameters by maximizing a composite likelihood function which measures the difference between the expected SFS and the observed data (Adams and Hudson, 2004). The calculation of composite likelihood typically ignores linkage between sites and assumes independence of each entry in the spectrum. The expected SFS is obtained by coalescent simulations in FASTSIMCOAL (Excoffier *et al.*, 2013, 2021; Nielsen, 2000) or diffusion approximation in DADI (Gutenkunst *et al.*, 2009; Kimura, 1964). Empirically, the joint SFS is rarely comprised of more than four populations due to its exponentially increasing dimensionality. Instead, approximation can be made to accommodate more species. For example, the composite likelihood can be replaced by the product of pairwise composite likelihoods, although this approach can complicate model comparisons (Excoffier *et al.*, 2013). Compared with DADI, the model of differential equations used in MOMENTS (Jouganous *et al.*, 2017; Ragsdale and Gravel, 2019, 2020) enables direct computation of

the frequency spectrum and is more tractable and scales up well to more populations.

1.3 Analysis Workflow

1.3.1 Multi-locus Data Preparation

Multi-locus data consist of multiple sequence alignments of short genomic segments, called *loci*. The use of multi-locus data is primarily determined by the model assumptions of MSC (Flouri *et al.*, 2018), which include the following: (1) free recombination between loci, (2) no recombination within a locus, (3) neutral evolution with no selection and (4) clock-like evolution. These assumptions imply certain desirable properties for datasets. For example, to satisfy assumptions (1) and (2), a common approach is to sample short genomic segments every few thousand base pairs along the genome. The size of a locus typically ranges between 100 and 2000 base pairs, with each locus separated by at least 2k to 10k base pairs. This ensures that loci are short enough for all sites within each locus to be fully linked, while being sufficiently far apart from each other to minimize linkage disequilibrium (LD) to assume independence of genealogical histories. The choice of distance between loci is essentially dependent on the system. For example, in humans, LD typically decays to background level within 20 – 100 kb (Jorde, 2000; Ribas *et al.*, 2008), whereas it decays more rapidly in chimpanzees.

Genome assemblies are constructed by mapping reads to the reference genome. Joint genotyping of multiple individuals is often performed at population or species level. Diploid (unphased) sequences are then built by extracting the genotype calls using coordinates obtained in the first step. Poorly sequenced regions (e.g., those with low depth or low genotyping scores) should be excluded or masked into *N*s in the sequences, and regions that are prone to alignment errors or potentially violate model assumptions (e.g., CpG islands, transposable elements, or highly repetitive sequences) should also be uniformly removed across all individuals, provided relevant genome annotations are available. In the context of low coverage data, excluding poorly sequenced individuals may be more favourable to maximize the number of loci retained for analysis. It is acceptable for different loci to have varying numbers of sequences, with some populations or species being present only in a subset

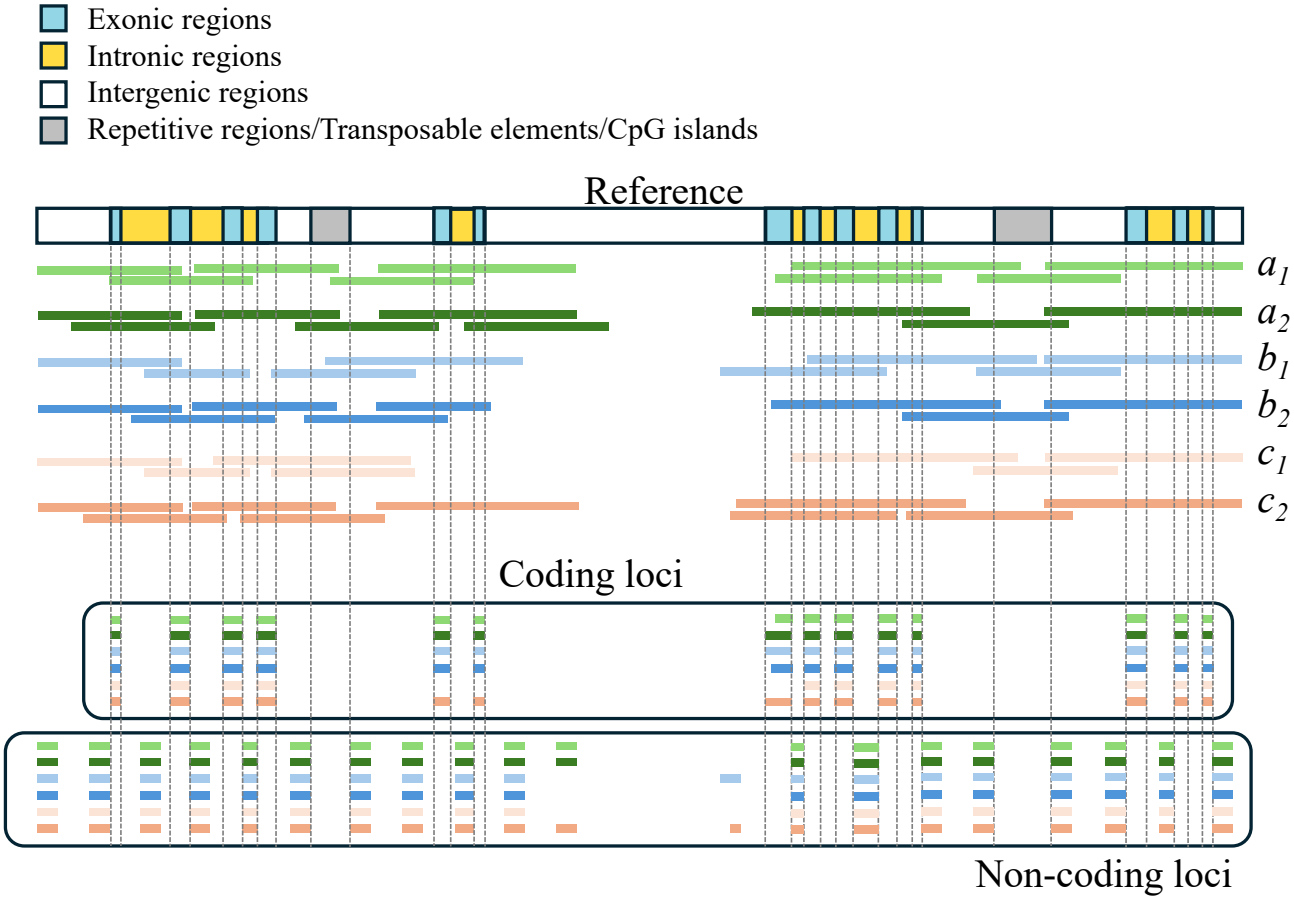

Figure 1.3: Generation of multi-locus data from sequencing reads. Two sets of loci are compiled. Coding loci include exons of genes, and non-coding loci are formed on intergenic and intronic regions. Short regions are sampled, which are separated by a certain distance in genome to achieve free recombination between them.

of loci. A minimum of one diploid or two haploid sequences are required. Otherwise, modern population sizes cannot be estimated with only a single haploid sequence per locus. Heterozygotes should be represented using IUPAC codes (e.g., R for A or G) in diploid sequences, and each can be phased into two haploid sequences using a proper haplotype estimation program (e.g., PHASE (Stephens and Donnelly, 2003; Stephens *et al.*, 2001)) or the analytical phasing implemented in BPP (Gronau *et al.*, 2011). Random phase resolution is documented to cause serious bias to parameter estimation of modern population sizes θ and introgression probability φ (Huang *et al.*, 2022b). As indicated by our simulation analysis, treating heterozygous sites as ambiguities can significantly mitigate the impact of genotyping errors in low-depth data (e.g., $< 10X$) and substantially improve the accuracy of species inference and parameter estimation for coalescent-based methods.

We may construct multiple datasets for genomic regions of different features, such as protein-

coding regions and non-protein-coding regions (i.e., intergenic and intronic regions). Exons evolve under the influence of selective forces, which may introduce a risk of bias in phylogenetic reconstruction. Technically, they are not qualified data for inferring species tree in the MSC model. However, in few studied systems, including the *Pan* genus in Chapter 4, the purifying selection acting on coding regions does not seem to alter the species tree but tends to yield more recent divergence times $\tau = T\mu$ (Shi and Yang, 2018; Thawornwattana *et al.*, 2022). It is possible when mutations are mostly neutral or deleterious. The uniform reduction of differences between species over these regions may not distort the topology of species tree. Nonetheless, as a precaution, it is advisable to treat exonic and noncoding regions as separate datasets in analysis.

1.3.2 Inference of Model of Gene Flow

Currently, obtaining a model of gene flow through systematic model construction is challenging. There are few methods that are able to perform model search using cross-model MCMC algorithms, such as SPECIESNETWORK in BEAST2 (Zhang *et al.*, 2018) and PHYLONET MCMC_SEQ (Wen and Nakhleh, 2018). In another Bayesian program IMA3 (Hey, 2010b; Hey *et al.*, 2018), uniform parameter priors are assigned to enable likelihood ratio tests of nested models (Hey and Nielsen, 2007; Nielsen and Wakeley, 2001). This can be used to exclude non-significant events in fitting of a saturated or near-saturated model of gene flow. None of the programs can handle datasets of > 500 loci due to poor mixing of MCMC (Jiao *et al.*, 2021). Here, we describe two heuristic approaches.

The first approach requires a stable species tree that is not misled by cross-species gene flow. In Chapter 2 of the thesis, we develop a Bayesian test of introgression (Ji *et al.*, 2023) that approximates the Bayes factor via the Savage-Dickey density ratio, and it can be used for comparing nested models in Bayesian framework. In this approach, introgression events are progressively added onto the species tree to construct a joint model with multiple introgression events. The Bayes test is applied to remove introgression events that are not supported in the model at each iteration. It is also advisable to start with a model specifying all possible introgression events and perform a stepwise subtraction process, as used in Chapter 3. For both strategies, pre-selecting a subset of plausible gene flow events beforehand can help reduce the required number of iterations. However, methods based on summary

statistics are not recommended for the selection as they are not capable of identifying gene flow between sister species. This approach should be considered for small to medium datasets involving up to few thousand loci, in which the data can be analysed as a whole or divided into a manageable number of subsets.

The other heuristic approach simply assembles gene flow signals on phylogeny. One common practice is to apply triplet-based methods to infer gene flow from all possible triplets (Suvorov *et al.*, 2022; Thawornwattana *et al.*, 2022). The detected signals are typically assumed in both directions and are subsequently translated into introgression edges between tips on the full phylogeny. The model is then revised to minimize the number of introgression events, which parsimoniously assumes that if gene flow is suggested between species A and most descendants of species B, it is considered to be a single introgression event between A and B. In Chapter 4, the idea is used to construct a migration model for the *Pan* genus. The efficiency of the approach is largely dependent on the applied triplet method, with which analysis is usually done relatively quickly. However, it does not compare models statistically when repositioning introgression edges on the phylogeny, not to mention the compromised power of summary methods in gene flow inference when their stringent model assumptions are violated (e.g., symmetrical population sizes, infinite sites model).

Chapter 2

Power of Bayesian and Heuristic Tests to Detect Cross-Species Introgression With Reference to Gene Flow in the *Tamias quadrivittatus* Group of North American Chipmunks

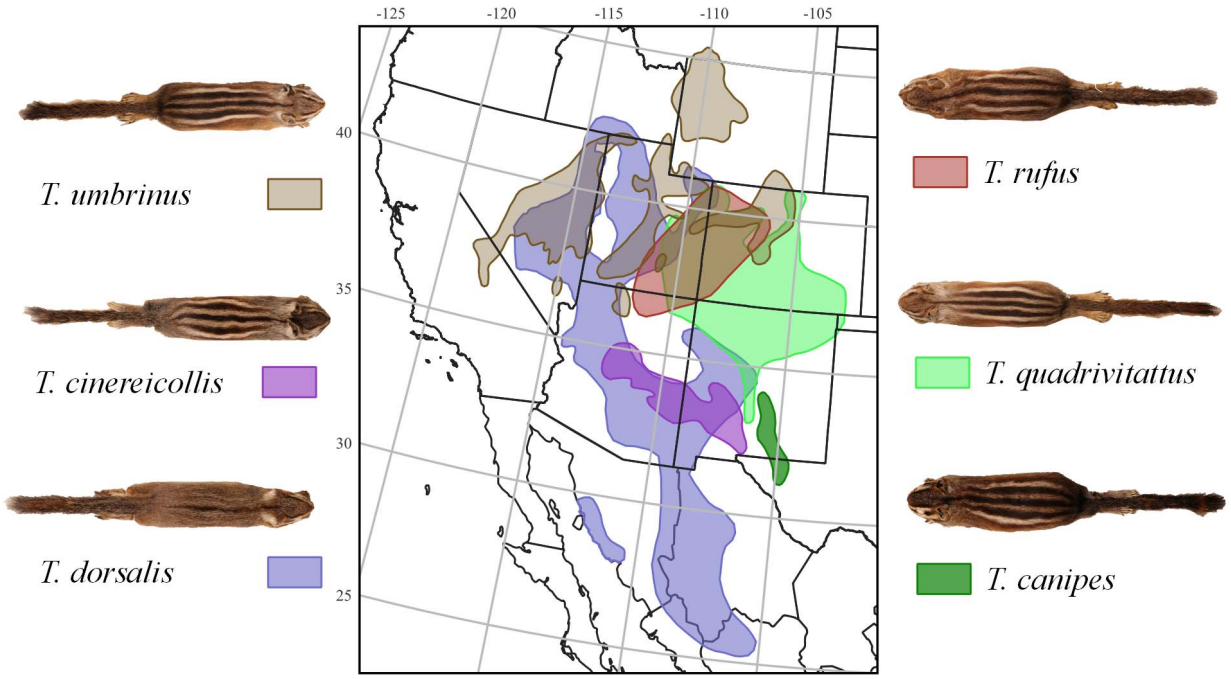

Figure 2.1: Geographic distributions of the six chipmunk species in the *Tamias quadrivittatus* group, based on data downloaded from the IUCN (`https://www.iucnredlist.org/`).

The *Tamias quadrivittatus* group of chipmunks currently consists of nine species that are distributed across the Great Basin along with the central and southern Rocky Mountains in North America (fig. 2.1). Previous work on *Tamias* has highlighted the importance of genital morphology, specifically the baculum (a bone found in the penis) in male chipmunks, as a reliable indicator of species limits (Patterson and Thaeler Jr, 1982; White, 2010). The biogeographic history of the group likely included large range fluctuations that have periodically resulted in isolation and secondary contact among species, which would have affected opportunities for hybridization and/or introgression (Good *et al.*, 2003). The current distributions of species in the group has extensive regions of overlap and

broad parapatry in ecological transition zones (fig. 2.1), with instances of both allopatry and parapatry, and the determinants of current distributions are thought to be related primarily to competitive exclusion and ecological preference (Brown, 1971; Heller, 1971; Root *et al.*, 2001). The system provides an exciting opportunity to investigate the effects of introgression on genetic variation within and between species.

Hybridization between chipmunk species has been widely reported based on discrepancies between mtDNA, nuclear DNA, and morphology (Good and Sullivan, 2001; Good *et al.*, 2003, 2008; Hird *et al.*, 2010). Work in the past decade has documented widespread mitochondrial introgression among species of the group (Reid *et al.*, 2012; Sarver *et al.*, 2017, 2021; Sullivan *et al.*, 2014), which is often asymmetrical, possibly due to bacular morphology, which has been identified in at least six species (Good *et al.*, 2003, 2008; Reid *et al.*, 2012; Sullivan *et al.*, 2014). Recent work on six species in the *T. quadrivittatus* group found that four of them exhibited clear evidence of introgressed mitochondrial DNA: *T. cinereicollis*, *T. dorsalis*, *T. quadrivittatus*, and *T. umbrinus* (table 2.1). The cliff chipmunk (*T. dorsalis*) was involved in local introgression with multiple other species, receiving mtDNA from whichever congeneric chipmunk it came into contact with. However, populations of *T. dorsalis* that are geographically isolated carry mtDNA haplotypes that are unique to the species (Sarver *et al.*, 2017; Sullivan *et al.*, 2014). Range overlap in transition zones plays an important role in mitochondrial introgression in *Tamias* (Bi *et al.*, 2019; Brown, 1971).

Table 2.1: Summary of evidence for mitochondrial introgression in the *T. quadrivittatus* group (Sullivan *et al.*, 2014)

Species	Region	Distribution	Introgression	Source
T. bulleri	M	Allopatric	No	
T. canipes (C)	GB/RM	Allopatric	No	
T. cinereicollis (I)	GB/RM	Parapatric	Yes	Not assignable
T. dorsalis (D)	GB/RM	Parapatric	Yes	C/U/Q/Not assignable
T. durangae	M	Allopatric	No	
T. palmeri	GB/RM	Allopatric	Untested	
T. quadrivittatus (Q)	GB/RM	Parapatric	Yes	Not assignable
T. rufus (R)	GB/RM	Allopatric	No	
T. umbrinus (U)	GB/RM	Parapatric	Yes	Not assignable

Note.— Geographic regions include Great Basin (GB), Rocky Mountains (RM), and Mexico (M). Single letter codes are for the six species included in the nuclear data analysis.

Sarver *et al.* (2021) used a targeted sequence-capture approach to sequence thousands of nuclear

loci (mostly genes or exons) to estimate the species phylogeny of the *T. quadrivittatus* group and to infer possible nuclear introgression. The program HYDE (Blischak *et al.*, 2018) was used to infer gene flow. Surprisingly, no significant evidence for gene flow involving the nuclear genome was detected between any species in the group, despite the evidence for widespread mitochondrial introgression. We note that HYDE, like the *D*-statistic, uses the four-taxon site-pattern counts pooled across the genome as data, and does not use information in the variation in genealogical history across the genome caused by the stochastic fluctuation of coalescent and introgression (Jiao *et al.*, 2021; Lohse and Frantz, 2014; Zhu and Yang, 2021). As a result, neither the *D*-statistic nor HYDE can detect gene flow between sister species or populations. Importantly, HYDE is designed to estimate the relative genetic contributions of the two parental species which hybridized to form a third species. When applied to detect other modes of gene flow, it makes restrictive assumptions about the direction of gene flow, and about species divergence times and population sizes that may be unrealistic (see fig. 2.7 below). The performance of HYDE when its model assumptions are violated is unexplored.

To examine whether the lack of evidence for nuclear introgression in the analysis of Sarver *et al.* (2021) may be due to the lack of power of HYDE, here we re-analyse the data of Sarver *et al.* (2021) using the BPP program (Flouri *et al.*, 2018, 2020), which includes a Bayesian implementation of the MSci model. Borrowing ideas from stepwise regression or Bayesian variable selection, we add introgression events sequentially onto the binary species tree to construct a joint MSci model with multiple introgression events. We develop a Bayesian test of introgression, calculating the Bayes factor for comparing the null model of no introgression against the alternative model of introgression via the Savage-Dickey density ratio (Dickey, 1971), using a Markov chain Monte Carlo (MCMC) sample under the MSci model. This may have a computational advantage over cross-model MCMC algorithms such as reversible jump MCMC (Green, 1995) or calculation of Bayes factors using thermodynamic integration (Gelman and Meng, 1998; Lartillot and Philippe, 2006). Our re-analysis revealed robust evidence for several ancient introgression events affecting the nuclear genome in the *Tamias* group, involving both sister species and nonsister species. We examine the model assumptions underlying HYDE and use computer simulation to demonstrate that the opposite conclusions reached in the two analyses may be explained by the lack of power of HYDE to detect gene flow. We then assess the

impact of ignoring introgression on estimation of population parameters, highlighting serious biases in species divergence time estimation when introgression exists and is ignored. Our results highlight the power of coalescent-based likelihood methods in the analysis of genomic datasets to infer the history of species divergence and gene flow.

2.1 Theory: Bayesian test of introgression

2.1.1 Bayes factor is given by the Savage-Dickey density ratio in comparisons of nested hypotheses

One can test for the presence of cross-species gene flow by comparing the introgression (MSci) model with the corresponding multispecies coalescent (MSC) model with no gene flow. The model of no gene flow (H_0) is a special case of the introgression model (H_1), with H_1 reducing to H_0 when the introgression probability is 0.

The commonly used device for Bayesian model comparison is the Bayes factor, which is the ratio of the marginal likelihood values under the two compared models. When the two models are nested, the Bayes factor is given by the Savage-Dickey density ratio (Dickey, 1971). In general, suppose we wish to compare the null model $H_0 : \phi = \phi_0$ against the alternative model $H_1 : \phi \neq \phi_0$, and suppose that both models have common (nuisance) parameters λ, while parameters ξ in H_1 become unidentifiable when $\phi = \phi_0$. The parameter vector is λ for H_0 and (ϕ, λ, ξ) for H_1. Given data x, let the likelihood be $L_0(\lambda)$ under H_0 and $L(\phi, \lambda, \xi) = p(x|\phi, \lambda, \xi)$ under H_1, with $L(\phi_0, \lambda, \xi) = L_0(\lambda)$ as the two models are nested. Let the prior be $\pi_0(\lambda)$ under H_0 and $\pi(\phi, \lambda, \xi) = \pi(\phi)\pi(\lambda|\phi)\pi(\xi|\phi, \lambda)$ under H_1. The Bayes factor in support of H_1 over H_0 is defined as

$$B_{10} = \frac{m}{m_0} = \frac{\iiint \pi(\phi, \lambda, \xi) L(\phi, \lambda, \xi)\, \mathrm{d}\phi \mathrm{d}\lambda \mathrm{d}\xi}{\int \pi_0(\lambda) L_0(\lambda)\, \mathrm{d}\lambda}, \tag{2.1}$$

where m_0 and m are the marginal likelihoods for the two models respectively.

Under the assumption that the priors on the common parameters (λ) agree between the two models, with

$$\pi(\lambda|\phi_0) = \pi_0(\lambda), \tag{2.2}$$

B_{10} can be expressed as the ratio of the prior and posterior densities for ϕ in H_1, both evaluated at the null value ϕ_0:

$$B_{10} = \frac{m}{m_0} = \frac{\pi(\phi_0)}{\pi(\phi_0|x)}, \tag{2.3}$$

where $\pi(\phi|x) = \iint \pi(\phi, \lambda, \xi|x)\,\mathrm{d}\xi\,\mathrm{d}\lambda$ is the marginal posterior density of ϕ.

Proof. Rewrite the prior $\pi_0(\lambda)$ and likelihood $L_0(\lambda)$ under H_0 as densities under H_1.

$$\begin{aligned} B_{10} &= \frac{m}{\int \pi_0(\lambda) L_0(\lambda)\,\mathrm{d}\lambda} \\ &= \frac{m}{\int \pi(\lambda|\phi_0) L_0(\lambda)\,\mathrm{d}\lambda} \\ &= \frac{m}{\int\int \frac{\pi(\phi_0,\lambda,\xi)}{\pi(\phi_0)} L(\phi_0,\lambda,\xi)\,\mathrm{d}\xi\,\mathrm{d}\lambda} \\ &= \frac{\pi(\phi_0)}{\int\int \frac{1}{m}\pi(\phi_0,\lambda,\xi) L(\phi_0,\lambda,\xi)\,\mathrm{d}\xi\,\mathrm{d}\lambda} \\ &= \frac{\pi(\phi_0)}{\iint \pi(\phi_0,\lambda,\xi|x)\,\mathrm{d}\xi\,\mathrm{d}\lambda} \\ &= \frac{\pi(\phi_0)}{\pi(\phi_0|x)}. \end{aligned} \tag{2.4}$$

Thus eq. 2.3 holds even if there exist nuisance parameters (λ) in both models, if the null values (ϕ_0) are at the boundary of the parameter space in H_1, and if some parameters in H_1 (ξ) become unidentifiable when the parameters of interest take the null values (when $\phi = \phi_0$). The proof above is more general than that given by Dickey (1971), which does not deal with the unidentifiability of ξ. Note that such irregular conditions cause considerable difficulties for likelihood ratio test (LRT), leading to unknown null distributions for the test statistic (e.g., Self and Liang, 1987). It is interesting that they do not cause any difficulty for the Bayesian test.

If the condition on the priors (eq. 2.2) does not hold, a correction factor may be applied (Verdinelli and Wasserman, 1995). This is not needed in our application.

2.1.2 Calculation of the Savage-Dickey density ratio

The prior density $\pi(\phi_0)$ of eq. 2.3 is typically available analytically. The posterior density $\pi(\phi_0|x)$ can be estimated using a kernel density smoothing procedure using the MCMC sample under H_1

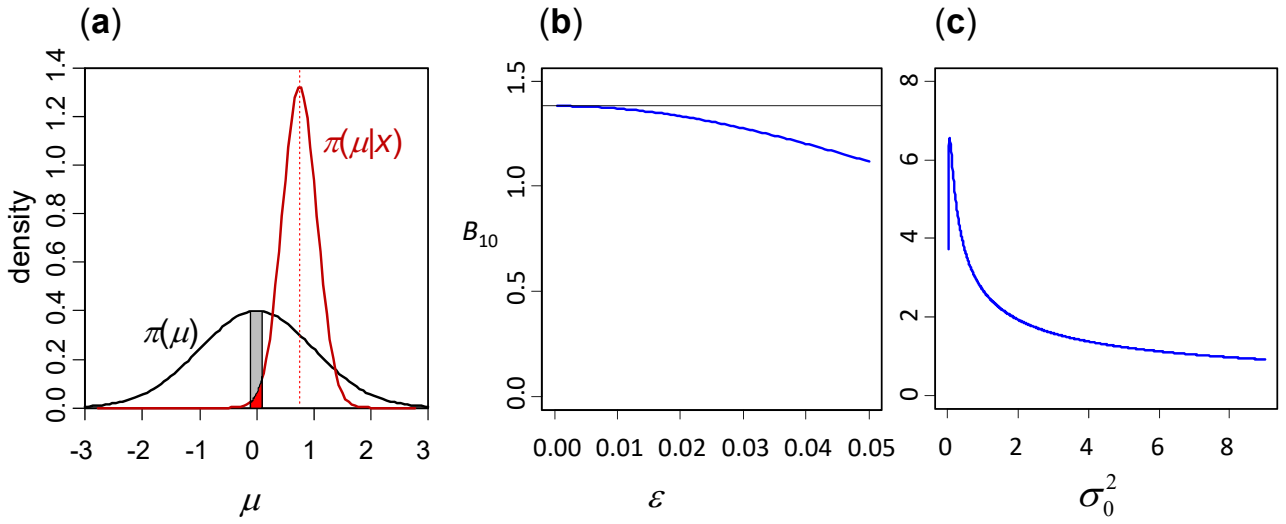

Figure 2.2: (**a**) Bayes factor expressed as the Savage-Dickey density ratio in the test of the null hypothesis $H_0 : \mu = 0$ against the alternative hypothesis $H_1 : \mu \neq 0$, using a data sample from $\mathbb{N}(\mu, 1)$. The black and red curves represent the prior and posterior densities for μ in H_1, and the small interval (of width ε) in the parameter space for H_1 is the null interval $\emptyset$ (or interval of null effects), representing H_0. The prior and posterior probabilities over the null interval (the gray and red areas) depend on the interval width (ε), but when $\varepsilon \to 0$, their ratio converges to the Bayes factor $B_{10} = \frac{\pi(\mu_0)}{\pi(\mu_0|x)}$. If the area of null effects shrinks greatly when we move from the prior to the posterior, the data contain strong evidence against H_0. (**b**) Approximate Bayes factor $B_{10,\varepsilon} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(\emptyset|x)}$ (eq. 2.8) plotted against ε for a dataset of size $n = 100$ with the sample mean $\bar{x} = 0.258$. The prior is $\mu \sim \mathbb{N}(0, \sigma_0^2)$ with $\sigma_0 = 2$ (twice the sampling standard deviation). When $\varepsilon \to 0$, $B_{10} = 1.381$. (**c**) Bayes factor (eqs. 2.1 or 2.13) plotted against the prior variance σ_0^2 for the same dataset showing the sensitivity of B_{10} to the prior on the parameter of interest (μ). Note that in this dataset (with $\sqrt{n}|\bar{x}| = 2.58$) H_0 is rejected by the LRT with p-value 1%.

(Silverman, 1986). This means that calculation of B_{10} using eq. 2.3 requires running the MCMC under H_1 only and no cross-model algorithms such as reverse-jump MCMC (Green, 1995) are needed. Note that within-model MCMC typically has better mixing properties than cross-model algorithms (Yang, 2014, pp. 247-260).

Suppose $(\phi^{(1)}, \phi^{(2)}, \cdots, \phi^{(N)})$ are an MCMC sample from the posterior $\pi(\phi|x)$. These are the ϕ values sampled during the MCMC, with the values for other parameters (λ and ξ) simply ignored. The kernel density estimator at the point ϕ_0 is

$$\hat{\pi}(\phi_0|x) = \frac{1}{Nh} \sum_{i=1}^{N} K\Big(\frac{\phi_0 - \phi^{(i)}}{h}\Big), \tag{2.5}$$

where $K(\cdot)$ is the kernel smoothing function and h is the smoothing parameter or window width. A

good choice of h is

$$h = 0.9 \cdot \min\left(\text{SD}, \frac{\text{inter-quartile range}}{1.34}\right) \times N^{-\frac{1}{5}} \tag{2.6}$$

(Silverman, 1986, eq. 3.30-3.31, p.47). The kernel function K is typically symmetrical around 0, with points further away from ϕ_0 make less contribution to the density at ϕ_0. For example, the Gaussian kernel is given as

$$K(t) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}. \tag{2.7}$$

However, this approach may be awkward to apply if the prior or posterior density at the null value, $\pi(\phi_0)$ or $\pi(\phi_0|x)$, is 0 or ∞. In this chapter, we use a more intuitive way of deriving the Savage-Dickey density ratio of eq. 2.3, which also provides an approach to its calculation. This treats the problem of testing as a problem of estimation, and assesses how likely the parameter of interest (ϕ) differs from the null value (ϕ_0). Define a null region or region of null effects, $\text{ø} : |\phi - \phi_0| < \varepsilon$, inside which ϕ is very close to ϕ_0. The null region is a small part of the parameter space for H_1 that represents H_0 (fig. 2.2). We then define a Bayes factor to represent the evidence for H_1

$$B_{10,\varepsilon} = \frac{1-\mathbb{P}(\text{ø}|x)}{\mathbb{P}(\text{ø}|x)} \Big/ \frac{1-\mathbb{P}(\text{ø})}{\mathbb{P}(\text{ø})} \approx \frac{\mathbb{P}(\text{ø})}{\mathbb{P}(\text{ø}|x)}, \tag{2.8}$$

as $1-\mathbb{P}(\text{ø}) \approx 1$ and $1-\mathbb{P}(\text{ø}|x) \approx 1$ for small ε. When $\varepsilon \to 0$, $\mathbb{P}(\text{ø}) \to \pi(\phi_0)\Delta$ and $\mathbb{P}(\text{ø}|x) \to \pi(\phi_0|x)\Delta$, where the differential Δ is the size of the null region, so that $B_{10,\varepsilon} \to \frac{\pi(\phi_0)}{\pi(\phi_0|x)}$, as in eq. 2.3. Thus the same conclusion is reached whether the problem is considered a testing problem (eqs. 2.1 or 2.3) or an estimation problem (eq. 2.8).

The approach is illustrated in figure 2.2 using the simple problem of testing $H_0 : \mu = 0$ against $H_1 : \mu \neq 0$ using a sample of size n from $\mathbb{N}(\mu, 1)$. The data are summarized as the sample mean $|\bar{x}|$. We assign the prior $\mu \sim \mathbb{N}(0, \sigma_0^2)$ under H_1. The posterior is then $\mu|x \sim \mathbb{N}(\mu_1, \sigma_1^2)$, with $\mu_1 = \frac{n\bar{x}}{n+1/\sigma_0^2}$ and $\frac{1}{\sigma_1^2} = n + \frac{1}{\sigma_0^2}$. The prior and posterior probabilities of the null interval are $\mathbb{P}(\text{ø}) = \mathbb{P}\{|\mu| < \varepsilon\} = 1 - 2\phi\left(-\frac{\varepsilon}{\sigma_0}\right) \approx \pi(\mu_0)\Delta$ and $\mathbb{P}(\text{ø}|x) = \phi\left(\frac{\varepsilon-\mu_1}{\sigma_1}\right) - \phi\left(\frac{-\varepsilon-\mu_1}{\sigma_1}\right) \approx \pi(\mu_0|x)\Delta$, with the differential to be the width of the null interval, $\Delta = 2\varepsilon$.

The above theory applies generally to Bayesian testing of nested hypotheses. Examples include comparison of different species delimitation models (e.g., one-species versus two-species models)

(Yang and Rannala, 2010) and test of migration between species (e.g., two species with and without migration) (Nielsen and Wakeley, 2001).

2.1.3 Test of introgression

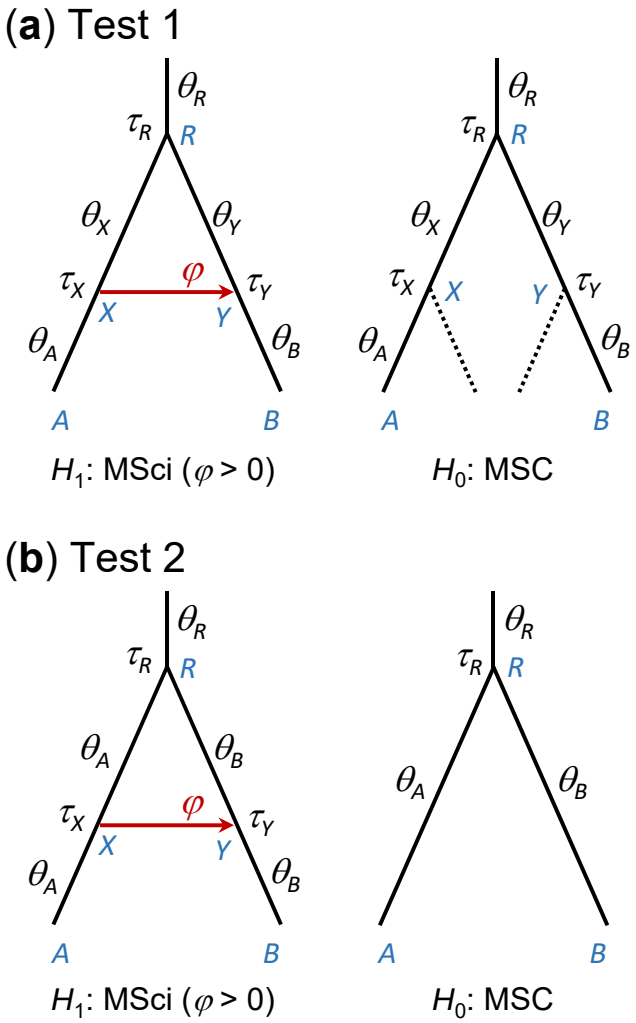

Figure 2.3: Parameters in the alternative and null hypotheses in two Bayesian tests of introgression (i.e., test of $H_0 : \varphi = 0$ against $H_1 : \varphi > 0$). The parameter of interest is the introgression probability φ. In test 1 (**a**), the shared parameters are $\lambda = (\tau_R, \tau_X = \tau_Y, \theta_A, \theta_B, \theta_R, \theta_X, \theta_Y)$. In test 2 (**b**), the shared parameters are $\lambda = (\tau_R, \theta_A, \theta_B, \theta_R)$ while $\xi = (\tau_X = \tau_Y)$ in H_1 becomes unidentifiable at the null value $\varphi_0 = 0$. Here only the two species involved in introgression are shown. Including other species on the species tree adds the same set of parameters to the null and alternative hypotheses.

When we use the Savage-Dickey density ratio (eq. 2.3) to test introgression, the nuisance parameters include species divergence times (τ) and population sizes (θ) on the species tree. Since we use the same priors on τ and θ in models with and without introgresion, independent of the introgression

probabilities (φ), the assumption of eq. 2.2 holds. We consider two tests with different assumptions about the population size parameters (fig. 2.3). In test 1, the MSci model assigns different θ parameters on the two segments of a branch broken by an introgression event; for example, in figure 2.3**a** branch RA is broken into two branches RX and XA and assigned θ_X and θ_A, respectively. The null model of no gene flow will have two θ parameters for the branch as well. Such a model can be implemented in BPP by including ghost species in the MSC model from which no sequences are sampled (fig. 2.3**a**). In the second test, the MSci model assigns the same θ parameter for a branch on the species tree before and after an introgression event (which can be specified using the control variable `thetamodel = linked-msci` in BPP) (fig. 2.3**b**). When the introgression probability takes the null value (0) in H_1, the introgression time τ_X becomes unidentifiable. The proof of eq. 2.4 applies to both scenarios. In this study, we used test 1. Note that calculating the Bayes factor using the Savage-Dickey density ratio (eqs. 2.3 or 2.8) requires an MCMC sample from H_1 and does not require any analysis or MCMC run under H_0.

In our BPP analysis, the introgression probability φ is assigned a beta prior beta(a,b), and the null hypothesis corresponds to $\varphi_0 = 0$ in H_1. Let the null region be $\emptyset : \varphi < \varepsilon$. Then $\mathbb{P}(\emptyset) = \mathbb{P}(\varphi < \varepsilon)$ in eq. 2.8 is given by the cumulative distribution function (CDF) for beta(a,b), while $\mathbb{P}(\emptyset|x)$ is simply the proportion of the sampled φ values that are $< \varepsilon$. Intuitively, the null region $\emptyset : \varphi < \varepsilon$ in H_1 represents absence of introgression (as the introgression probability φ is negligibly small), $\frac{1-\mathbb{P}(\emptyset)}{\mathbb{P}(\emptyset)}$ is the prior odds in favor of gene flow, while $\frac{1-\mathbb{P}(\emptyset|x)}{\mathbb{P}(\emptyset|x)}$ is the posterior odds, and B_{10} measures the change in the odds in favour of gene flow when we move from the prior to the posterior. We used $\varepsilon = 0.01$ and confirm that use of $\varepsilon = 0.001$ gave very similar results. A cut-off of 20 for B_{10} may be considered strong evidence in support of H_1 (corresponding to 95% posterior for H_1 if the prior model probabilities for H_0 and H_1 are $\frac{1}{2}$ each), while 100 means extremely strong evidence (corresponding to 99% posterior for H_1).

2.2 Materials and Methods

2.2.1 Chipmunk genomic data

The dataset, generated and analysed by Sarver *et al.* (2021), includes 1060 nuclear loci from six chipmunk species: *T. rufus* (R), *T. canipes* (C), *T. cinereicollis* (I), *T. umbrinus* (U), *T. quadrivittatus* (Q) and *T. dorsalis* (D) (with 5, 5, 9, 10, 11, 11 individuals, respectively), as well as the outgroup *T. striatus* (3 individuals). We included all individuals whether or not their mtDNA was likely to be introgressed. Due to lack of a reference genome, Sarver *et al.* (2021) assembled genomic loci (targeted genes or exons) into contigs using an approach called Assembly by Reduced Complexity (ARC). Filters were then applied to remove missing data (contigs not present across all individuals) and sequences with likely assembly errors. The procedure generated a dataset of 1060 loci (1060 ARC contigs, Sarver *et al.*, 2021), with sequence length ranging from 14 to 1026 bp among loci and the number of variable sites from 0.33% to 15.2%.

High-quality heterozygous sites in the data, as identified by high mapping quality and depth of coverage, are represented using IUPAC ambiguity codes. They are accommodated using the analytical integration algorithm implemented in BPP (Flouri *et al.*, 2018; Gronau *et al.*, 2011). This takes the unphased genotype sequences as data and averages over all possible heterozygote phase resolutions, using their relative likelihoods based on the sequence alignment at the locus as weights (Huang *et al.*, 2022b).

2.2.2 Species tree estimation for the *T. quadrivittatus* group

We used BPP version 4 (Flouri *et al.*, 2018; Rannala and Yang, 2017) to estimate the species tree under the MSC model without gene flow. This is the A01 analysis (`speciesdelimitation=0`, `speciestree=1`) (Yang, 2015).

We assigned inverse-gamma (IG) priors to parameters in the MSC model: $\theta \sim$ IG(3, 0.002) with mean 0.001 for population size parameters and $\tau_0 \sim$ IG(3, 0.01) with mean 0.005 for the age of the root. The shape parameter $\alpha = 3$ means that those priors are diffuse, while the prior means are based on estimates from preliminary runs. Note that both θ and τ are measured in the expected number

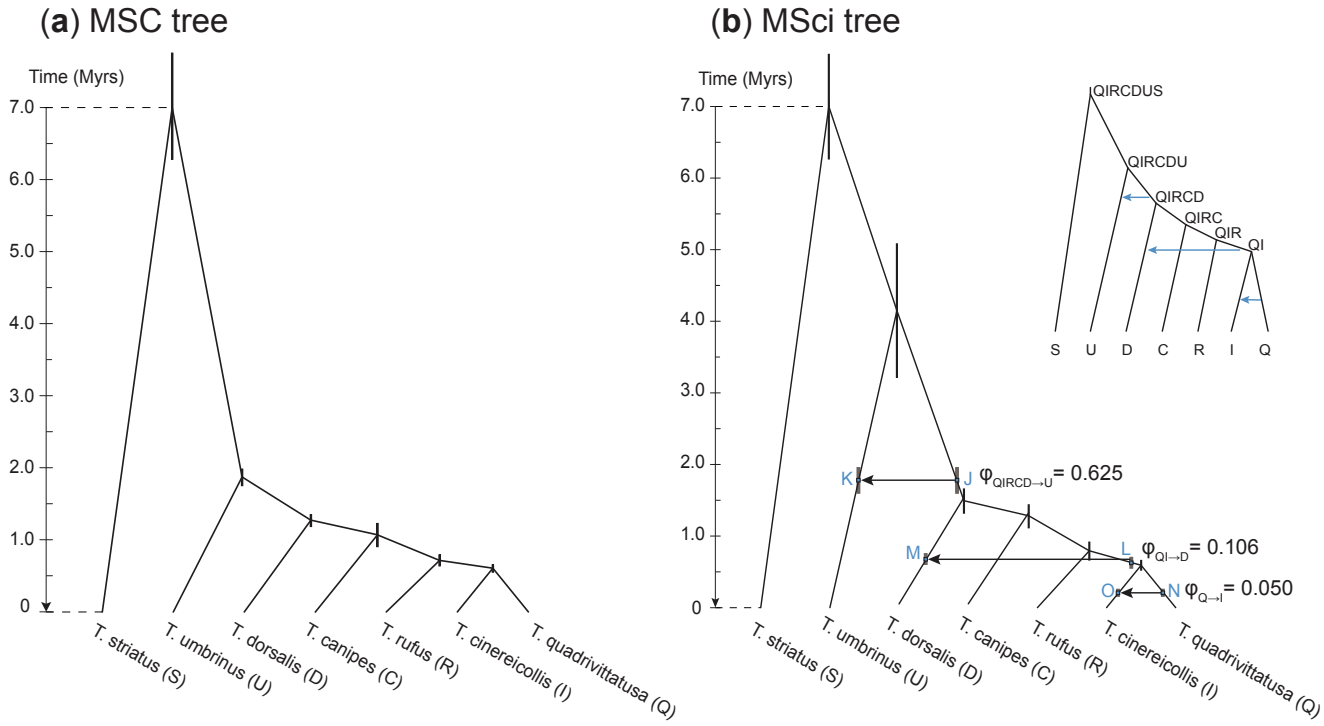

Figure 2.4: (**a**) Species tree for the *T. quadrivittatus* group with *T. striatus* used as the outgroup. Branch lengths represent the posterior means of divergence times (τ) estimated from BPP analysis of the full data of 1060 loci under the MSC model with no gene flow, with node bars indicating the 95% HPD intervals. A minimum divergence time of 7 Myrs for the outgroup *T. striatus* is used to convert the τ estimates into absolute times. **b**) The joint introgression model constructed in this study with three unidirectional introgression events, showing parameter estimates from BPP analysis of the full data of 1060 loci. Nodes created by introgression events are labeled, with the labels used to identify parameters in table S2.3. The MSci model includes 6 species divergence times and 3 introgression times (τ), 19 population size parameters (θ), and 3 introgression probabilities (φ).

of mutations per site. The inverse gamma is a conjugate prior for θ and allows the θ parameters to be integrated out analytically, leading to a reduction of parameter space and improved mixing of the MCMC algorithm. We conducted 10 replicate MCMC runs, using different starting species trees. Each run generated 2×10^5 samples, with a sampling frequency of 2 iterations, after a burn-in of 16,000 iterations. Each run took about 70 hours using one thread on a server with Intel Xeon Gold 6154 3.0GHz processors. Convergence was confirmed by consistency between runs. All runs converged to the same species tree (fig. 2.4**a**), with $\sim$ 100% posterior probability, which had the same topology as the tree inferred by Sarver *et al.* (2021).

2.2.3 Stepwise construction of the introgression model

As the species tree is well supported, apparently unaffected by cross-species introgression, we used the species tree to build an introgression model with multiple introgression events. Our procedure is similar to stepwise regression, the step-by-step method for constructing a regression model that involves adding or removing explanatory variables based on a criterion such as an F-test or t-test.

Our procedure has two stages. In the first stage, we used BPP to fit a number of introgression models, each with only one introgression event, and rank candidate introgression events by their strength (indicated by the introgression probability φ). The analyses of Sarver *et al.* (2021) suggest that mitochondrial introgression affected mostly four species: *T. umbrinus* (U), *T. dorsalis* (D), *T. quadrivittatus* (Q) and *T. cinereicollis* (I). We considered introgression events involving all possible pairs among those four species, as well as another species, QI, the common ancestor of *T. cinereicollis* and *T. quadrivittatus* (fig. 2.4**a**). The dataset of 1060 loci was analysed under an MSci model with only one introgression event, estimating the introgression probability (φ) and introgression time (τ). We assign the same inverse-gamma priors on θ and τ as above, and beta$(1,1)$ or $\mathbb{U}(0,1)$ for the introgression probability φ. Two replicate runs were conducted for each analysis to confirm consistency between runs, and MCMC samples from the two runs were then combined to produce posterior estimates of parameters. This analysis provides a ranking of the introgression events by the introgression probability. We calculated the Bayes factor for testing $H_0 : \varphi_0 = 0$ given by the Savage-Dickey density ratio (eq. 2.3), using the null interval $\emptyset = (0, 0.01)$ (eq. 2.8); use of $(0, 0.001)$ produced virtually identical results. Only introgresssion events with $B_{10} \geq 20$ were considered further.

In the second stage, we added introgression events onto the binary species tree (fig. 2.4**a**) sequentially in the order of decreasing strength (introgression probability). To reduce the computational cost and to examine the robustness of the analysis, this step was applied to two subsets of the 1060 loci: the first half and the second half, each of 530 loci. The priors used for population sizes and root age were as above. With multiple introgression events in the model, we extended the MCMC runs to be k-times as long if the model involved k introgression events. Three replicate runs were performed to check consistency between runs. Samples from the replicate runs were then combined to produce posterior summaries. At each step, the added introgression event was retained if it met the same cutoff

as above in either of the two data subsets.

Our procedure produced a joint introgression model with three unidirectional introgression events. The joint model was then applied to the full dataset of 1060 loci to estimate the population parameters including introgression probabilities, introgression times, species divergence times, and population sizes (fig. 2.4**b**), using the same prior settings. We conducted 3 replicate runs, using a burn-in of 50,000 iterations and then taking 10^6 samples, sampling every 2 iterations. Each run took 200 hrs.

2.3 Results

2.3.1 Species tree estimation for the *T. quadrivittatus* group

We analysed the full data of 1060 loci under the MSC model without gene flow to estimate the species tree. The ten replicate runs using different starting species trees converged to the same maximum *a posteriori* probability (MAP) tree, with posterior probability $\sim 100\%$ (fig. 2.4**a**). Sarver *et al.* (2021) recovered the same species tree topology in their analysis of the same data using ASTRAL (Mirarab and Warnow, 2015) and SVDQUARTETS (Chifman and Kubatko, 2014), although with weaker support for some nodes, e.g., concerning the placement of *T. rufus*. The differences in support may be due to the fact that ASTRAL and SVDQUARTETS use summaries of the multilocus sequence data that are not sufficient statistics, and are thus less efficient than the full likelihood method implemented in BPP (Xu and Yang, 2016; Zhu and Yang, 2021).

2.3.2 Stepwise construction of the introgression model

In the first stage of our procedure, we fitted introgression models, each involving one introgression event, using the full dataset of 1060 loci. We considered introgression events between every contemporary pair of the five species: *T. cinereicollis* (I), *T. dorsalis* (D), *T. quadrivittatus* (Q), and *T. umbrinus* (U), and the ancestral species QI (fig. 2.4**a**). Introgression events that passed our cutoff ($B_{10} \geq 20$) are listed in table 2.2. Introgression from QI into D had the highest probability, $> 10\%$, while six more events had $\varphi > 5\%$: Q→D, D→QI, QI→U, I→D, Q→I, and I→Q. We note that introgressions between Q and I, and between QI and D, was significant in both directions and the

Table 2.2: Posterior means and 95% HPD CIs (in parentheses) for introgression probability (φ) and introgression time (τ) in the separate introgression analysis

	Introgression	φ	τ ($\times 10^{-3}$)	B_{10}
*	QIRCD → U	0.6215 (0.3907, 0.8243)	0.896 (0.784, 1.004)	∞
*	QI → D	0.1187 (0.0866, 0.1499)	0.337 (0.311, 0.367)	∞
	Q → D	0.0779 (0.0509, 0.1026)	0.297 (0.253, 0.328)	∞
	D → QI	0.0707 (0.0384, 0.1058)	0.337 (0.302, 0.366)	∞
	QI → U	0.0624 (0.0269, 0.1020)	0.408 (0.353, 0.457)	21.27
	I → D	0.0579 (0.0332, 0.0862)	0.265 (0.217, 0.318)	∞
*	Q → I	0.0568 (0.0315, 0.0750)	0.098 (0.073, 0.121)	∞
	I → Q	0.0533 (0.0153, 0.0969)	0.111 (0.077, 0.156)	∞
	D → U	0.0214 (0.0022, 0.0483)	0.276 (0.178, 0.474)	0.04
	Q → U	0.0198 (0.0037, 0.0389)	0.296 (0.209, 0.367)	0.05
	D → I	0.0180 (0.0092, 0.0275)	0.155 (0.123, 0.192)	0.39
	D → Q	0.0177 (0.0058, 0.0315)	0.184 (0.117, 0.347)	0.10
	U → QI	0.0097 (0.0022, 0.0181)	0.371 (0.322, 0.410)	0.01
	I → U	0.0069 (0.0015, 0.0136)	0.158 (0.098, 0.223)	0.01
	U → D	0.0066 (0.0024, 0.0112)	0.235 (0.176, 0.300)	0.01
	U → Q	0.0061 (0.0008, 0.0127)	0.200 (0.119, 0.294)	0.01
	U → I	0.0037 (0.0009, 0.0071)	0.147 (0.090, 0.207)	0.01

Note.— The species tree of figure 2.4**a** is used, with a single introgression event assumed in each analysis. The full dataset of 1060 loci is analysed using BPP to estimate the introgression probability (φ) and the introgression time (τ), together with the species divergence times (τ) and population sizes (θ) on the species tree. Introgression events with $B_{10} < 20$ (D → U and below) are not considered further in the stepwise approach of constructing the joint introgression model. The three introgression events that are selected in the joint introgression model are marked with asterisks. Bayes factor $B_{10} = \infty$ occurs if all φ values in the MCMC sample are $> \varepsilon = 1\%$.

estimated introgressions times were close (table 2.2). We thus replaced the two unidirectional introgression events by one bidirectional introgression in further analyses (model D in Flouri *et al.*, 2020).

The time of QI→U introgression was estimated to be 0.000408, very close to the species divergence time at node QIR (0.000417) (fig. 2.4**a**), suggesting that the introgression was probably a more ancient event. Note that if an introgression event is assigned incorrectly to a daughter branch to the lineage truly involved in introgression, one would expect the estimated introgression time to collapse onto the species divergence time. We thus attempted to place the introgression onto more ancient ancestral branches on the species tree (fig. 2.4**a**) and finally identified the lineage involved in introgression to be the ancestral species QIRCD. The QIRCD→U introgression had an estimated time that was away from the species divergence times, and the estimated introgression probability (62%) was

the highest (table 2.2).

In the second stage, we added introgression events identified in table 2.2 onto the binary species tree of figure 2.4**a**, in the order of their introgression probabilities (table S2.1). This was applied to two data subsets (the full data split into two halves). While our procedure allows introgression events already in the model to drop out when new introgressions are added to the model, this did not happen in the analysis of the *Tamias* dataset. Instead the most important introgression events identified in stage 1 remained to be most important in the joint introgression models constructed in stage 2. Note that multiple introgression events may not be independent. An introgression event significant in stage 1 may not be significant anymore when other introgression events are already included in the model. For example, when the QI→D introgression was already included in the model, none of the introgressions Q→D, D→QI, I→D and I→Q was significant. Those introgressions may be expected to lead to similar features in the sequence data, such as reduced sequence divergences between Q or I and D. Similarly, introgression probability for an introgression event often became smaller when other introgressions were added in the model. However, the opposite may occur as well. For example, $\varphi_{QIRCD\to U}$ was estimated to be 54-63% when this was the only introgression assumed in the model, but increased to 59-69% when other introgression events were added in the model (table S2.1).

Results for the two data subsets were largely consistent, especially concerning introgression events with high introgression probabilities. We thus arrived at a joint introgression model with three unidirectional introgression events (fig. 2.4**b**, table S2.1).

We examined the impact of the prior for φ on the Bayesian test of introgression. We calculated the Bayes factor B_{10} using the full dataset of 1060 loci under the prior $\varphi \sim \text{beta}(\alpha,\beta)$, with $\alpha = 0.2, 1, 5$ and $\beta = 0.2, 1, 5$, generating nine prior settings (table S2.2). Note that beta(α,β) has the mean $\mathbb{E}(\varphi) = \frac{\alpha}{\alpha+\beta}$ and variance $\mathbb{V}(\varphi) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. In particular, the prior mean varied from 0.0385 for beta(0.2, 5) to 0.961 for beta(5, 0.2). The Bayes factor B_{10} was ∞ for all three introgression probabilities in the joint model, insensitive to the prior on φ (table S2.2).

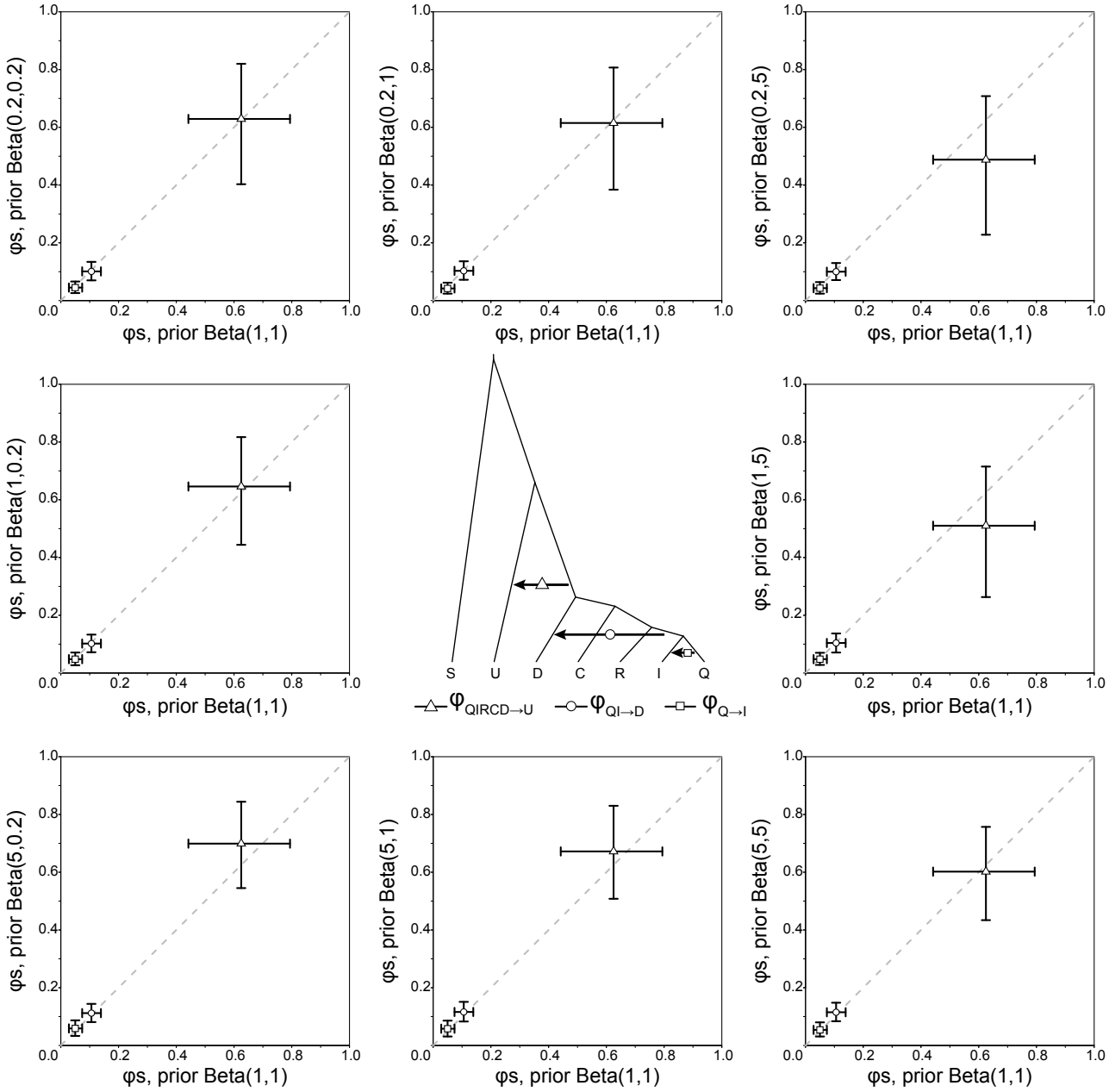

Figure 2.5: Posterior means and 95% HPD CIs for the three introgression probabilities (φ) obtained from BPP analyses of the full data of 1060 loci using different beta priors, $\varphi \sim$beta(α, β).

2.3.3 Estimation of introgression probabilities and species divergence/introgression times

Finally, we fitted the joint introgression model of figure 2.4**b** to the full data of 1060 loci, as well as the two halves, with parameter estimates shown in table S2.3. The fitted model is very parameter-rich, partly as we assign different θ parameters for different branches on the species tree: for example, branch Q in figure 2.4**b** is broken into two segments by the introgression event, Q→I, which are assigned two independent θ parameters. As a result, population sizes for ancestral species tend to be poorly estimated, especially for those populations with a very short time duration. These patterns are consistent with simulation studies that examine the information content in multi-locus datasets (Huang *et al.*, 2020).

The estimated introgression probabilities from the full data are 0.625 with the 95% highest probability density (HPD) credibility interval (CI) to be (0.442, 0.794) for $\varphi_{QIRCD\to U}$, 0.106 (0.074, 0.139) for $\varphi_{QI\to D}$, and 0.050 (0.028, 0.074) for $\varphi_{Q\to I}$. The introgression probability $\varphi_{QIRCD\to U}$ involved considerable uncertainty, with a large CI, possibly because the introgression is ancient and is between sister species, making it hard to estimate its strength, so that the dataset of 1060 loci may be too small.

We evaluated the impact of the prior for φ on parameter estimation in the analysis of the full dataset, using $\alpha = 0.2, 1, 5$ and $\beta = 0.2, 1, 5$ in the prior $\varphi \sim \text{beta}(\alpha, \beta)$ (fig. 2.5). The prior had some effects on $\varphi_{QIRCD\to U}$, with the prior mean being more important than the prior variance. Under beta(0.2, 5) with the prior mean 0.0385, the posterior mean was lower, and the CI wider. Under beta(5, 0.2) with the prior mean 0.961, the posterior mean was higher, and the CI narrower. However, the posterior CIs overlapped considerably among the different priors, and overall the impact of the prior for φ on the estimate of $\varphi_{QIRCD\to U}$ was minor. Estimates of $\varphi_{QI\to D}$ and $\varphi_{Q\to I}$ were insensitive to the prior used (fig. 2.5).

Accommodating gene flow in the model had significant impacts on estimation of the time of divergence between species involved in gene flow (figs. 2.4 & 2.6). While estimates of times for the recent divergences ($\tau_{QI}, \tau_{QIR}, \tau_{QIRC}$, and τ_{QIRCD}) were nearly identical between the MSC model ignoring gene flow and the MSci model incorporating gene flow, the estimated age of the *T. quadrivittatus* clade (τ_{QIRCDU}) was much greater under MSci than under MSC (fig. 2.6). This can be explained by the fact that the MSC model ignored the QIRCD→U introgression, which had introgression proba-

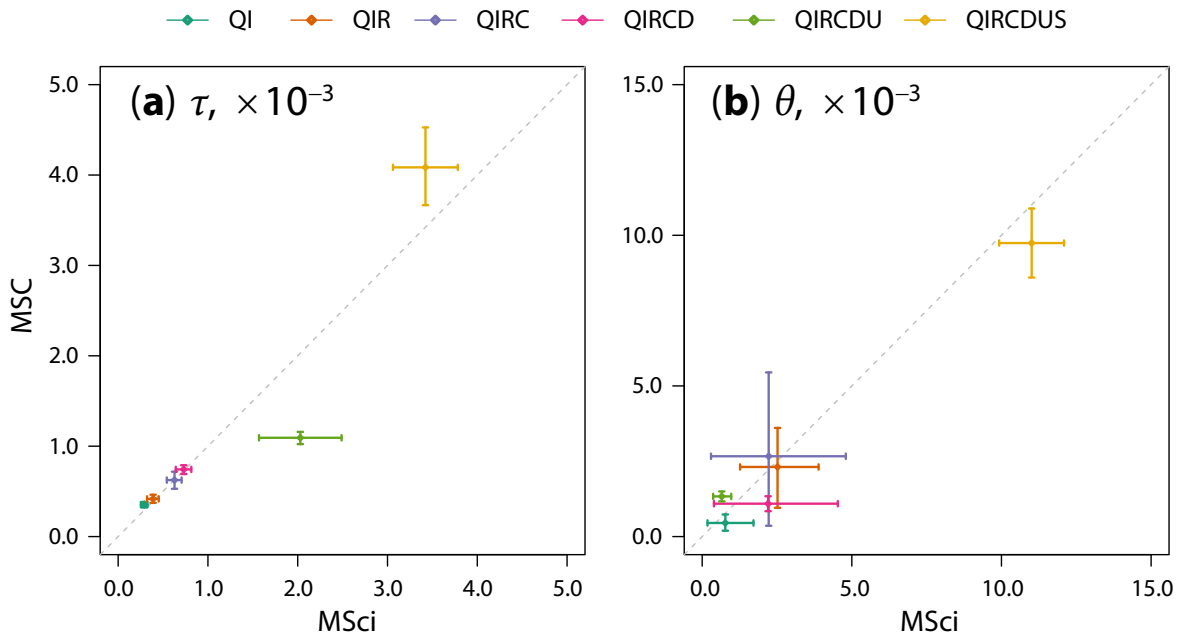

Figure 2.6: Scatterplot of posterior means and 95% HPD CIs (**a**) for the six species divergence times (τ) and (**b**) for the six ancestral population sizes (θ) in the MSC and MSci models of figure 2.4 obtained from BPP analyses of the full data of 1060 loci. Note that both τ and θ are measured in the expected number of mutations per site.

bility 62.5%. Note that sequence divergence between any pair of species X and Y has to be older than species divergence ($t_{XY} > \tau_{XY}$), and as a result, the minimum (rather than average) sequence divergence dominates the estimate of species divergence time. If gene flow is present between species and is ignored in the model, the reduced sequence divergence due to gene flow will be misinterpreted as recent species divergence, leading to underestimation of species divergence time. This effect has been noted in previous simulations (Leaché *et al.*, 2014).

The estimated age of the root of the species tree ($\tau_{QIRCDUS}$) was slightly smaller under MSci than under MSC. However, $\tau_{QIRCDUS}$ is negatively correlated with the population size ($\theta_{QIRCDUS}$) so that both parameters have large uncertainties (Burgess and Yang, 2008).

Sullivan *et al.* (2014, fig. 1) used the minimum divergence time of 7 Ma for the outgroup species *T. striatus*, based on fossil teeth thought to belong to *Tamias* found in the late Miocene, reported in Dalquest *et al.* (1996), to date the *T. quadrivittatus* clade to 1.8 Ma in a maximum-likelihood concatenation analysis of four nuclear genes, and to 1.2 Ma (with 95% CI 0.6–2.2) in a *BEAST (Heled and Drummond, 2010) analysis of the same data. Concatenation analysis is known to be biased as it does not accommodate the stochastic variation of gene tree topologies and divergence

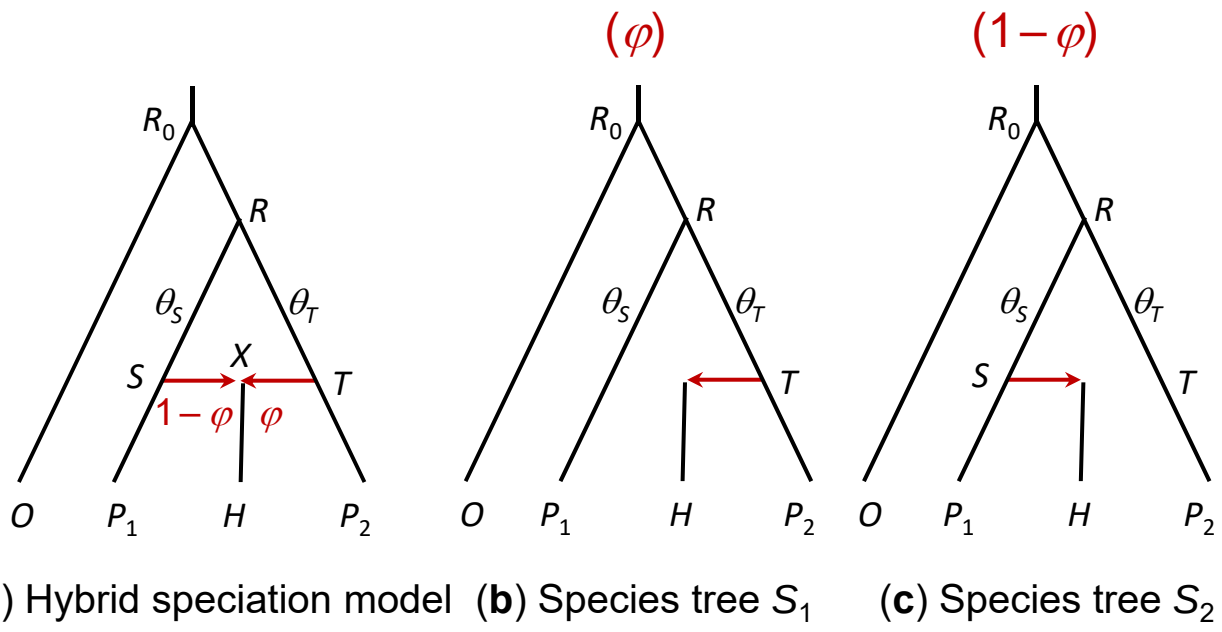

Figure 2.7: (**a**) HYDE assumes a hybrid-speciation model with the additional assumption of equal population sizes, or a symmetrical inflow model, with $\tau_S = \tau_T$ and $\theta_S = \theta_T$ (Blischak *et al.*, 2018). (**b, c**) Two parental species trees S_1 and S_2 induced by the hybridization model of (**a**). Site patterns are a mixture over the two species trees.

times among loci due to the coalescent process (Ogilvie *et al.*, 2017). We used the same calibration to rescale the estimates of τ under the MSC and MSci models (fig. 2.4). The minimum age for the *T. quadrivittatus* clade was 1.9 Ma (with 95% HPD CI to be 1.8–2.0) under the MSC model, comparable to the *BEAST estimate under the same model (fig. 2.4**a**). Under the MSci model, the estimated minimum age was 4.1 Ma (with CI be 3.2–5.1) (fig. 2.4**b**), much older than the estimates under the MSC model without gene flow. Note that here the CIs accommodate the uncertainty due to finite amounts of sequence data but not uncertainties in the fossil calibration.

2.3.4 Model assumptions underlying HYDE

Whereas the analyses of nuclear data by Sarver *et al.* (2021) using HYDE detected no significant signal of introgression at all, our BPP analyses of the same data revealed strong evidence of multiple introgression events, involving both sister and non-sister species (fig. 2.4**b**). To understand the opposing conclusions reached in the two analyses, here we examine the model assumptions underlying HYDE. We then use simulation to compare the performance of HYDE and BPP under conditions that are representative of the *Tamias* data but may violate the assumptions of HYDE.

HYDE was developed under the hybrid-speciation model of figure 2.7**a**, with $\tau_S = \tau_X = \tau_T$, and $\theta_S = \theta_T$ (Blischak *et al.*, 2018). Formulated for quartet data, with one sequence from each of the four species, it uses the counts or frequencies of three parsimony-informative site patterns: $iijj$, $ijji$, $ijij$, to estimate the genetic contributions of the two parental species to the hybrid species: φ and $1-\varphi$. Here pattern $ijkl$ means a site with nucleotides i, j, k, l in O, P_1, H, P_2, respectively (fig. 2.7**a**). Under this model, the probabilities of gene trees and site patterns are both given by a mixture over the two binary species trees S_1 and S_2 (called *parental species trees*), with mixing probabilities φ and $1-\varphi$ (fig. 2.7**b**&**c**). Given species tree S_1, the matching pattern $iijj$ has a larger probability (say, a) than the other two mismatching patterns (each with probability b, say, with $b < a$). Given species tree S_2, the matching pattern $ijji$ has probability a while the two mismatching patterns have b each. The symmetry assumptions ($\tau_S = \tau_T$ and $\theta_S = \theta_T$) ensure that a, b for tree S_1 are equal to a, b for S_2. By averaging over the two species trees, the site pattern probabilities under the hybridization model are given as

$$\begin{aligned} p_{iijj} &= \varphi a + (1-\varphi) b \\ p_{ijij} &= \varphi b + (1-\varphi) b = b \\ p_{ijji} &= \varphi b + (1-\varphi) a. \end{aligned} \tag{2.9}$$

Setting those probabilities to the observed frequencies ($\hat{p}$) and eliminating a and b from the system of equations gives the estimate

$$\hat{\varphi} = \frac{\hat{p}_{iijj} - \hat{p}_{ijij}}{\hat{p}_{iijj} - 2\hat{p}_{ijij} + \hat{p}_{ijji}}, \tag{2.10}$$

This is eq. 3 in Blischak *et al.* (2018), although the derivation here is simpler than that of Kubatko and Chifman (2019). Note that the theory works if $\tau_S = \tau_T > \tau_X$ and $\theta_S = \theta_T$, so that the method may be used under model A of Flouri *et al.* (2020, fig. 1) with the symmetry assumption. The null hypothesis of no hybridization/introgression ($H_0 : \varphi = 0$) can be tested by applying a normal approximation to the site-pattern counts (Kubatko and Chifman, 2019).

To see which of the two assumptions ($\tau_S = \tau_T$ and $\theta_S = \theta_T$) has more impact, note that a change in τ is comparable with the same amount of change in $\frac{2}{\theta}$. Coalescent may occur in population RS (if the H sequence takes the left parental path in the model of fig. 2.7a), at the rate $\frac{2}{\theta_S}$ over time

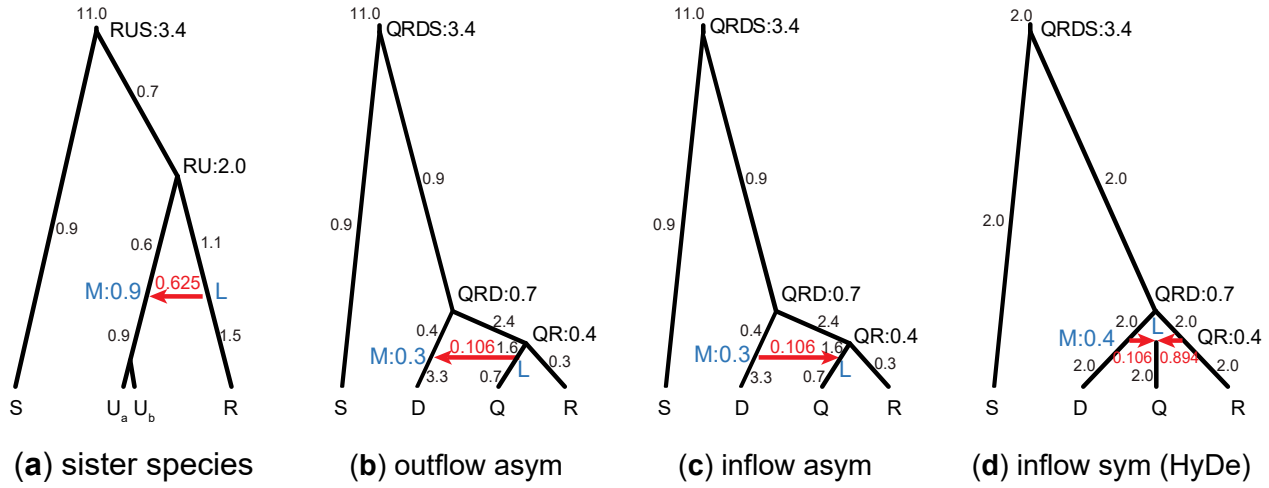

Figure 2.8: Introgression models (species trees with introgression) used for simulating data to evaluate the performance of HYDE and BPP. **(a)** Species tree for three species (R, U and S) with $R \rightarrow U$ introgression at the rate of $\varphi = 0.625$, and with S to be the outgroup, based on BPP estimates from the *Tamias* data (fig. 2.4**b**, table S2.3). Population sizes (θ) are next to the branches and species divergence times (τ) are next to the nodes. Two sequences are sampled from species U. When the data are analysed using HYDE, either Ua or Ub is specified as the hybrid lineage. (**b**) Outflow model for three species (D, Q, R), with S to be the outgroup, with introgression from Q to D at the rate $\varphi = 0.106$ (table S2.3). (**c**) Inflow asymmetrical model for three species, with asymmetrical divergence times and population sizes. (**d**) Inflow symmetrical model for three species, with $\tau_M = \tau_{QR}$ and $\theta_M = \theta_{QR}$ (see fig. 2.7**a**). Note that only model (**d**) matches the assumption of HYDE.

period $\tau_R - \tau_S$, and it may occur in population RT (if the H sequence takes the right parental path), at the rate $\frac{2}{\theta_T}$ over time period $\tau_R - \tau_T$. If $\frac{2(\tau_R - \tau_S)}{\theta_S} = \frac{2(\tau_R - \tau_T)}{\theta_T}$, the probability of coalescent (given that two sequences enter populations S or T) will be the same in the two populations. However, the probabilities of the site patterns depend on the time of coalescent as well as its occurrence. Thus for eq. 2.10 to be valid, both the rates and the times have to be identical: $\tau_S = \tau_T$ and $\theta_S = \theta_T$.

Note that HYDE or the D-statistic cannot be used to infer gene flow between sister lineages. One might think that HYDE or D could be applicable if two sequences were sampled from the recipient lineage to form a quartet. However this is not the case. With ancient introgression, the two sequences from the same lineage are interchangeable and have the same average genomic distance to the outgroup sequence. Suppose P_1 and H in figure 2.7**a** are two sequences from the same lineage. Then site patterns $iijj$ and $ijij$ will have the same probability even if $\varphi > 0$.

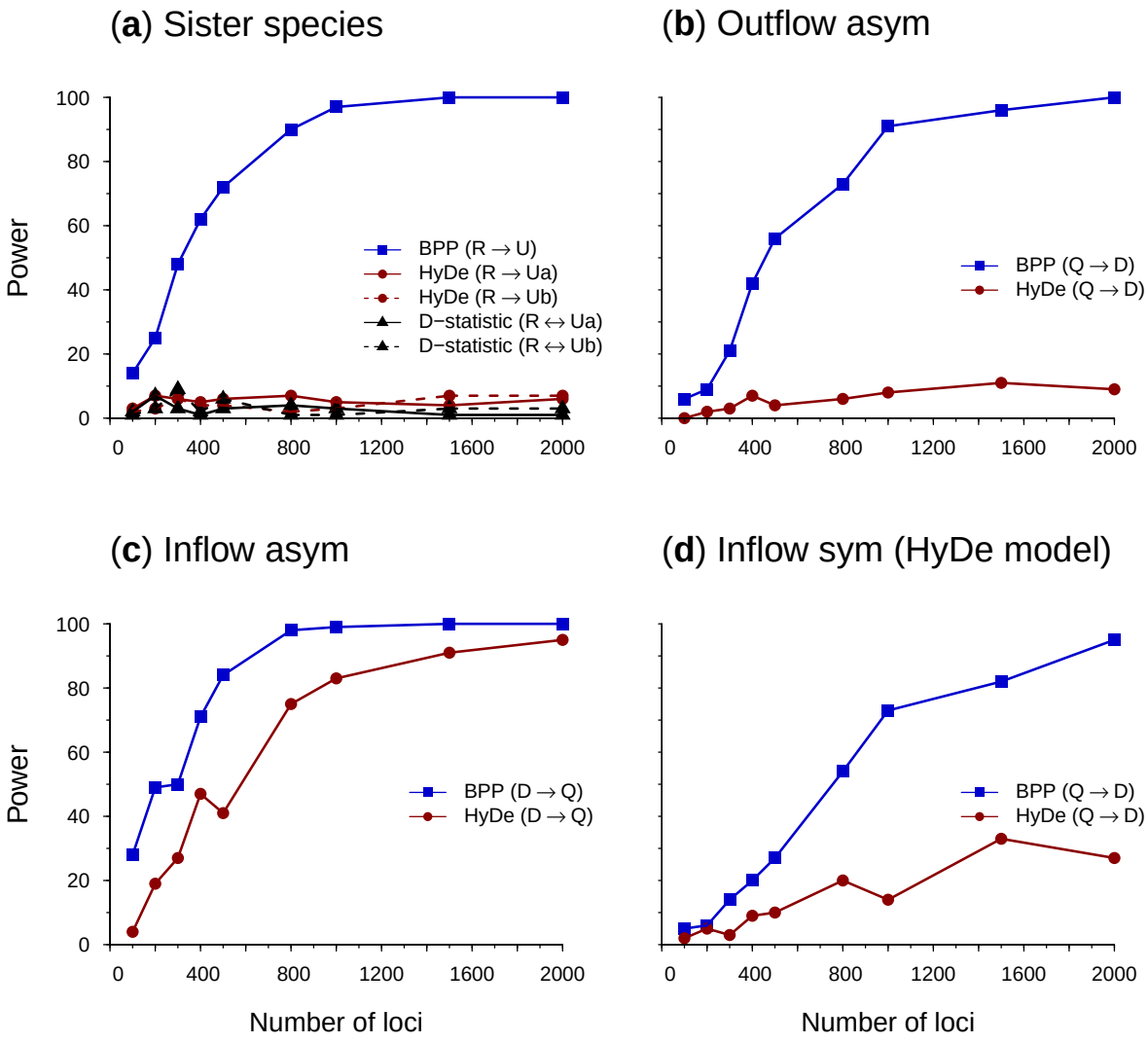

Figure 2.9: Power of detecting gene flow by HYDE and BPP in 100 replicate datasets simulated under the models of figure 2.8.

2.3.5 Simulations to examine the performance of HYDE

Our examination of assumptions underlying HYDE suggests that HYDE may not be suitable for testing gene flow in the *Tamias* data. The strongest introgression in the *Tamias* data detected using BPP was between sister species, with $\varphi_{QIRCD\to U} = 0.625$ (fig. 2.4**b**). This is unidentifiable by HYDE. The next introgression involved outflow with $\varphi_{QI\to D} = 0.106$, whereas HYDE assumes inflow. The third introgression was again between sister species, with $\varphi_{Q\to I} = 0.050$. To verify those expectations and to explore the performance of HYDE and BPP under different scenarios of gene flow, we conducted simulations using four different model settings (fig. 2.8**a**-**d**), based on parameter estimates obtained

from the *Tamias* data (fig. 2.4**b**, table S2.3). Gene trees and sequence alignments at multiple loci were generated using the `simulate` option of BPP. HYDE analysis was conducted using PAUP (Swofford, 2003). The data were also analysed using BPP. The results are summarized in figure 2.9.

Model a (fig. 2.8**a**) assumes gene flow between sister lineages, based on the introgression event from QIRCD→U in the *Tamias* data (fig. 2.4**b**). It was suggested that by including multiple sequences from the recipient lineage, HYDE or the *D*-statistic might be used to detect gene flow between sister lineages. We used species R and U, with introgression rate $\varphi_{R\to U} = 0.625$, including two sequences (Ua and Ub) from the recipient species U, while S was used as the outgroup. The divergence times (τ) and population sizes (θ) were based on the real data (table S2.3). When multiple branches in the full tree (fig. 2.4b) were merged into one branch in the tree of figure 2.8**a**, θ for the merged branch was calculated as a weighted average, with the branch lengths as weights. As our objective in this case was to confirm the lack of power of HYDE (and the *D*-statistic), we simulated large datasets, each with $L = 8000$ loci. The sequence length was 500 sites, and the number of replicates was 100. When the data were analysed using HYDE and the *D*-statistic, the quartet tree (((Ua, Ub), R), S) was used, with Ua or Ub labelled the 'hybrid' lineage. The same data were analysed using BPP under the MSci model with three species (fig. 2.8**a**).

As expected, HYDE and the *D*-statistic had no power to detect gene flow between sister lineages: indeed, the power of HYDE and *D* was not higher than the significant level (fig. 2.9, table S2.4). Note that a test that ignores data and produces 5% positives at random will have 5% of power. Also HYDE did not produce reliable estimates of φ; in about half of the datasets, the estimate was outside the range $(0, 1)$.

Model b (fig. 2.8**b**) was based on the next strongest introgression in the *Tamias* data, with $\varphi_{QI\to D} = 0.106$ (fig. 2.4**b**). We used species D, Q, R, with S as the outgroup. This is a case of outflow, when gene flow from an ingroup species Q to a more distant species D. Our examination of the assumptions made by HYDE suggests that HYDE can be used to detect inflow, but not outflow. We generated datasets of various sizes with $L = 500, 2000$ or 8000 loci. The other settings were the same as for model a. When the data were analysed using HYDE, Q was designated the 'hybrid' lineage while R and D were the two parents. HYDE performed poorly (fig. 2.9**b**), with very low power and frequent

invalid estimates of φ (table S2.5).

Model c (fig. 2.8**c**) was the same as model b but the direction of gene flow was reversed. The model was then a case of inflow, as assumed by HYDE. However, species divergence times and population sizes did not satisfy the symmetry requirements of HYDE (in other words, $\tau_M \neq \tau_{QR}$ and $\theta_M \neq \theta_{QR}$). In this case, HYDE had considerable power in detecting gene flow (fig. 2.9**c**). However, the estimates of φ by HYDE involved large biases, apparently converging to ≈ 0.32 when the true value was 0.106 (table S2.5). This positive bias is apparently because coalescent occurs at a higher rate or over longer time period on the M branch than on the QR branch in figure 2.8**c**, with $\frac{\tau_{QRD}-\tau_M}{\theta_M} > \frac{\tau_{QRD}-\tau_{QR}}{\theta_{QR}}$. In the opposite case, the bias should be negative.

Model d (fig. 2.8**d**) was the same as model c with inflow but in addition we enforced the symmetry assumptions, so that species Q was a hybrid species formed by hybridization between D and R. This is the hybrid speciation model assumed by HYDE, and the method performed well (fig. 2.9**d**). Its power was lower than that for BPP, as expected from statistical theory, but improved with the increase of data, rising from 10% at $L = 500$ loci to 90% at 8000 loci. The parameter estimate appeared to be consistent, converging to the correct value (0.106) when the number of loci increased, and there were not many invalid estimates (table S2.5). Those results are consistent with previous simulations, which evaluated the performance of HYDE when all its assumptions were met and found the method to perform well (Blischak *et al.*, 2018; Flouri *et al.*, 2020).

In summary, our simulations suggest that it is important to apply HYDE to detect the correct mode of gene flow (that is, gene flow between non-sister lineages, and inflow instead of outflow) (fig. 2.8d). Furthermore, the symmetry assumptions are important for HYDE to produce reliable estimates of introgression probability. When all model assumptions are met, HYDE performed well. However, HYDE had no power to detect gene flow between sister lineages, and very low power to detect outflow.

In all four models (fig. 2.8**a**-**d**), the Bayesian test using BPP had good power (fig. 2.9, tables S2.4&S2.5). Furthermore, the posterior means and 95% HPD CIs for parameters in the introgression models b-d were well-behaved (fig. 2.10). While HYDE can estimate only two parameters from the site-pattern counts (the internal branch length in coalescent units on the species tree and the introgres-

sion probability), the BPP analysis of the same data estimates all parameters in the model. The species divergence/introgression times were all well estimated with small CIs (fig. 2.10). The introgression probability was accurately estimated with narrow CIs when ≥ 500 loci were used. Population size parameters for short branches were poorly estimated due to lack of coalescent events in those populations.

Table 2.3: False positive rate of BPP and HYDE tests and average estimates of introgression probability in 100 simulated replicates

	BPP			HYDE			
# loci	Error Rate ($\alpha = 1\%$)	Error Rate ($\alpha = 5\%$)	$\hat{\varphi}\pm$SD	Error Rate ($\alpha = 1\%$)	Error Rate ($\alpha = 5\%$)	$\hat{\varphi}\pm$SD	Proportion of invalid estimates
Inflow asym (fig. 2.8**c**)							
500	0%	0%	0.019 ± 0.011	1%	7%	0.140 ± 0.108	52%
2000	0%	0%	0.009 ± 0.004	5%	13%	0.094 ± 0.061	52%
8000	0%	0%	0.004 ± 0.002	2%	7%	0.038 ± 0.032	51%
Inflow sym (fig. 2.8**d**, HYDE model)							
500	0%	0%	0.032 ± 0.016	0%	3%	0.064 ± 0.048	49%
2000	0%	0%	0.014 ± 0.006	1%	2%	0.039 ± 0.029	55%
8000	0%	0%	0.006 ± 0.003	0%	3%	0.022 ± 0.016	49%

Note.— Data were simulated using the species trees of figure 2.8**c**-**d** but with $\varphi = 0$.

We also examined the false positive rate (type-I error rate) of the HYDE and Bayesian tests, by simulating data using the inflow-asym (fig. 2.8**c**) and inflow-sym (fig. 2.8**d**) models but with $\varphi = 0$ fixed so that there was no introgression in the true model. The results are summarized in table 2.3. Under the inflow-asym model, HYDE had higher false positive rate than the nominal significant level. For example, at the 5% significance level, the false positive rate was 7%, 13%, and 7% in datasets of 500, 2000, and 8000 loci, respectively. The high rate may be explained by the violation of the symmetry assumptions for HYDE. Under the inflow-sym model (or the HYDE model), the rate was 3%, 2%, and 3%, all within the allowed 5% (table 2.3). Thus HYDE performed well when its assumptions were met and had elevated false positives when the assumptions were violated. In all settings, the false positive rate of the Bayesian test was estimated to be $\sim 0\%$. This is consistent with the expectation that the Bayesian test may be more conservative (with lower false positive rate and lower power) than the LRT (see discussions later).

Finally, to assess the information content in datasets of the size of the *Tamias* data, we used

Table 2.4: LRT and Bayesian tests in the normal example in two datasets

Data	LRT	Bayesian test		
$\sqrt{n}\lvert\bar{x}\rvert$	p-value	Prior	B_{10}	$\mathbb{P}(H_1\lvert x)$
1.96	$p = 0.05$	$\sigma_0 = 1$	0.359	0.264
1.96	$p = 0.05$	$\sigma_0 = 2$	0.262	0.208
1.96	$p = 0.05$	$\sigma_0 = 10$	0.120	0.107
2.58	$p = 0.01$	$\sigma_0 = 1$	0.408	0.290
2.58	$p = 0.01$	$\sigma_0 = 2$	0.300	0.230
2.58	$p = 0.01$	$\sigma_0 = 10$	0.138	0.122

Note.— The Bayes factor B_{10} is calculated assuming data size $n = 100$ in eq. 2.13, while the posterior model probability is given by eq. 2.14. Note that the p-value for the LRT is 5% (or 1%) in the dataset with $\sqrt{n}\lvert\bar{x}\rvert = 1.96$ (or 2.58).

parameter estimates from the full dataset (fig. 2.4**b**, table S2.3) to simulate two datasets of the same size as the original, with 5, 5, 9, 10, 11, 11, 3 unphased sequences per locus for species R, C, I, U, Q, D and S, respectively. The sequence length was 200 sites. We analysed the datasets under the same MSci model of figure 2.4**b** using BPP to estimate all parameters. The estimates from the two datasets were similar, so we present those from one of them in table S2.3. At this data size, BPP achieved relatively good precision and accuracy. The posterior means were close to the true values, and the CIs were also similar to those calculated from the real data. Similarly to analyses of the real data, divergence times and population sizes for modern species were well estimated, but ancestral population sizes, in particular those for populations of short time duration, were more poorly estimated.

2.4 Discussion

2.4.1 Criteria for testing gene flow

Hypothesis testing or model selection involves arbitrariness, and classical hypothesis testing and Bayesian model selection applied to the same data may produce strongly opposed conclusions, a situation known as Jeffreys's paradox (Jeffreys, 1939; Lindley, 1957). Furthermore, Bayesian model selection is known to be sensitive to priors on model parameters, especially on parameters that are not shared between the models under comparison. See Yang (2014, pp.194-7) for a discussion of those issues. Here we review different strategies for testing, using as example a simple problem

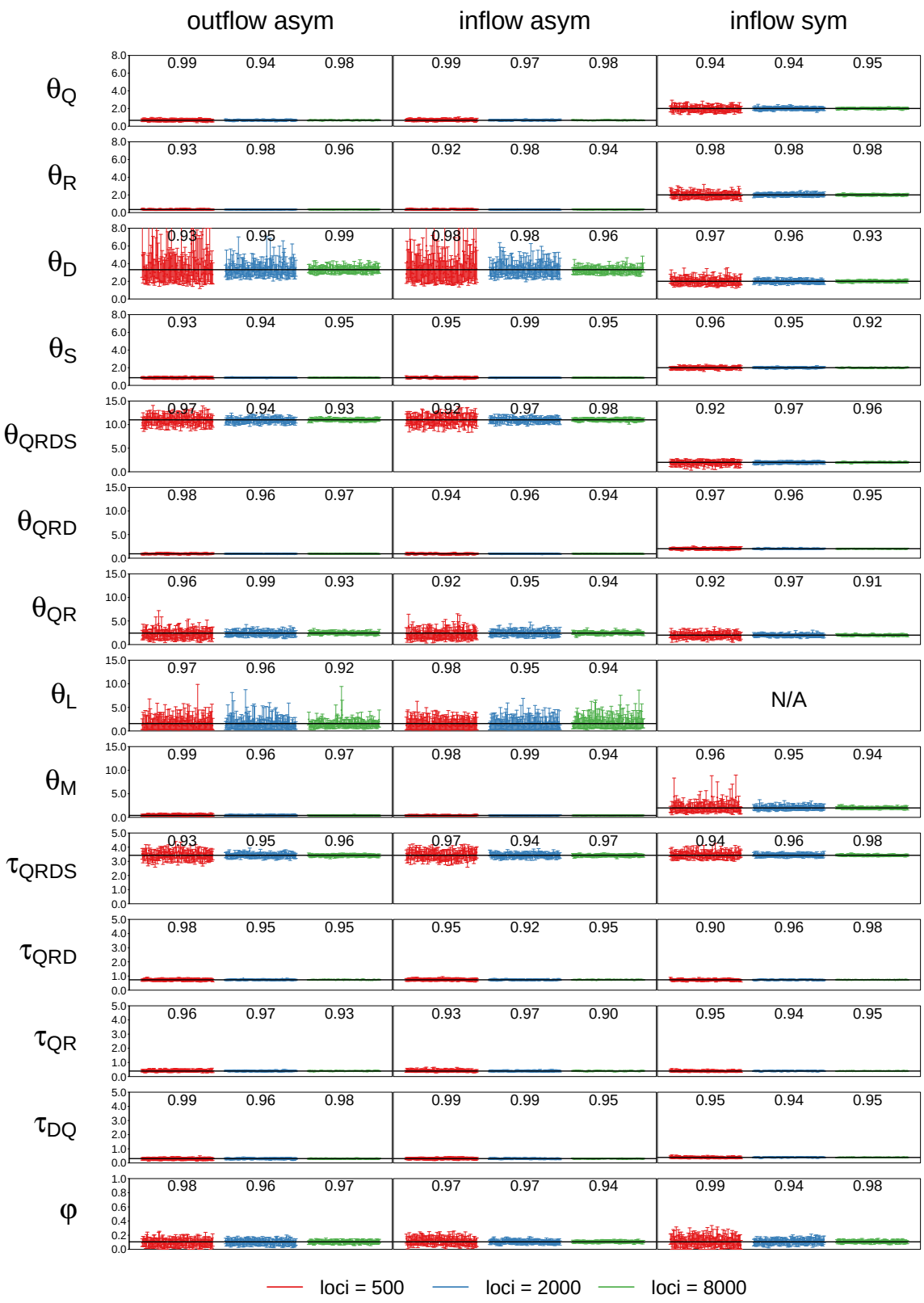

Figure 2.10: Posterior means and 95% HPD CIs for parameters in the three introgression models of figure 2.8: (**b**) outflow asym, (**c**) inflow asym and (**d**) inflow sym (HYDE model), in BPP analyses of 100 replicate datasets, each with 500, 2000, or 8000 loci. Note that in model (d) inflow sym, all populations had the same size (θ) although separate θ parameters were estimated for different populations when the data were analysed using BPP. Parameters τ and θ are multiplied by 10^3. The number above the CI bars is the coverage or the probability that the CI includes the true value.

of testing the null hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$, using a data sample, $x = \{x_1, x_2, \cdots, x_n\}$, from the normal distribution $\mathbb{N}(\mu, 1)$. We assume that a false positive error (of falsely rejecting H_0 when it is true) is more serious than a false negative error (of failing to reject H_0 when it is false). The data can be summarized as the sample mean $\bar{x}$, with the likelihood given by $\bar{x} \sim \mathbb{N}(0, 1/n)$ under H_0 and $\bar{x} \sim \mathbb{N}(\mu, 1/n)$ under H_1. Let $\phi(x; \mu, \sigma^2)$ be the probability density function (PDF) for $\mathbb{N}(\mu, \sigma^2)$ and $\Phi(\cdot)$ be the CDF for $\mathbb{N}(0, 1)$.

In hypothesis testing, the p-value can be calculated from the fact that under H_0, $\sqrt{n}|\bar{x}| \sim \mathbb{N}(0, 1)$ or $n|\bar{x}|^2 \sim \chi^2_1$. At the $\alpha = 5\%$ significance level, we reject H_0 if

$$2\Delta\ell = 2\log \frac{\phi(\bar{x}; \bar{x}, \frac{1}{n})}{\phi(\bar{x}; 0, \frac{1}{n})} = n|\bar{x}|^2 > \chi^2_{1,5\%} = 3.84. \tag{2.11}$$

Alternatively one may consider this as an estimation problem and construct a confidence interval (CI) for μ and reject H_0 if the CI excludes the null value 0. This is equivalent to the LRT.

In a Bayesian analysis, we consider two approaches. The first is to examine whether the posterior 95% credibility interval (CI) for μ under H_1 excludes the null value 0. We assign the prior $\mu \sim \mathbb{N}(0, \sigma_0^2)$ under H_1. The posterior is then $\mu|x \sim \mathbb{N}(\mu_1, \sigma_1^2)$, with mean $\mu_1 = \frac{n\bar{x}}{n+1/\sigma_0^2}$ and precision $\frac{1}{\sigma_1^2} = n + \frac{1}{\sigma_0^2}$. Here the reciprocal of variance is known as precision. The sample precision is n and the prior precision is $1/\sigma_0^2$, while the posterior precision is the sum of the two. The 95% CI for μ is given as $\mu_1 \pm 1.96\sigma_1$ so that the CI excludes 0 (in which case we reject H_0) if $|\mu_1| > 1.96\sigma_1$, or if

$$n|\bar{x}|^2 > 3.84\big[1 + 1/(n\sigma_0^2)\big]. \tag{2.12}$$

The second approach is to use the Bayes factor to compare the null and alternative hypotheses.

$$\begin{aligned} B_{10} &= \frac{\mathbb{P}(\bar{x}|H_1)}{\mathbb{P}(\bar{x}|H_0|} = \frac{\phi(\bar{x}; 0, \frac{1}{n} + \sigma_0^2)}{\phi(\bar{x}; 0, \frac{1}{n})} \\ &= \frac{1}{\sqrt{1 + n\sigma_0^2}} \cdot \exp\left\{ \frac{n\bar{x}^2}{2\big[1 + 1/(n\sigma_0^2)\big]} \right\}, \end{aligned} \tag{2.13}$$

(e.g., Yang, 2006, eq. 5.21).

The Bayes factor is closely related to (and 'calibrated' using) the posterior model probability. If the two models are assigned equal prior probabilities ($\pi_0 = \pi_1 = \frac{1}{2}$), the posterior model probability is

$$\mathbb{P}(H_1|x) = \frac{B_{10}}{1+B_{10}}, \tag{2.14}$$

so that a 95% cut-off on $\mathbb{P}(H_1|x)$ corresponds to $B_{10} = 19$, and H_0 is rejected based on the Bayes factor if and only if

$$n|\bar{x}|^2 > \log\left\{19\sqrt{1+n\sigma_0^2}\right\} \times 2\left[1+1/(n\sigma_0^2)\right]. \tag{2.15}$$

While the LRT (eq. 2.11) depends on $\sqrt{n}|\bar{x}|$ only, both the posterior CI (eq. 2.12) and the Bayes factor (eq. 2.15) depend in addition on $n\sigma_0^2$. Note that the three criteria (eqs. 2.11, 2.12, & 2.15) have the ordering

$$\begin{aligned} 3.84 &< 3.84\left[1+1/(n\sigma_0^2)\right] \\ &< \log\left\{19\sqrt{1+n\sigma_0^2}\right\} \times 2\left[1+1/(n\sigma_0^2)\right]. \end{aligned} \tag{2.16}$$

Thus the LRT has more power and higher false positive rate than the posterior CI while the Bayesian test based on the Bayes factor is the most conservative. The result reflects the general perception that the LRT tends to reject the null hypothesis and favour parameter-rich models too often, especially in large datasets. Note that if H_0 is true, the false positive rate of the LRT stays at 5% when the sample size $n \to \infty$, whereas in the Bayesian analysis, the true model H_0 will dominate, with $\mathbb{P}(H_0|x) \to 1$ and $B_{10} \to 0$ when $n \to \infty$.

Example calculations are given in table 2.4 for two datasets with $\sqrt{n}|\bar{x}| = 1.96$ or 2.58 and $n = 100$. In both datasets, H_0 is rejected by the LRT (at the 5% and 1% levels, respectively), but the Bayes factor and the posterior model probabilities favour H_0 over H_1, with $B_{10} < 1$ and $\mathbb{P}(H_1|x) < \frac{1}{2}$.

This analysis suggests that the difference in power between HYDE and BPP are due to the inefficient use of information in the data by HYDE, not to the different statistical philosophies. An LRT for testing introgression applied to the multilocus sequence alignments may be expected to have more power (and higher false positive rate) than the Bayesian test based on the Bayes factor.

2.4.2 The power of heuristic and likelihood methods to detect introgression

When applied to the *Tamias* dataset, HyDe and bpp produced opposite conclusions concerning gene flow. Our examination of the model assumptions for HyDe and our simulations suggest that this is because gene flow with the strongest signal in the *Tamias* group, either between sister species or involving outflow, may be of the wrong type or in the wrong direction for HyDe. Here we review and summarize the major issues with HyDe.

First, both HyDe and the *D*-statistic pool sites across loci when counting site patterns, so that the site-pattern counts are genome-wide averages. Cross-species gene flow creates genealogical variation across the genome, with the probabilistic distribution of the gene trees and coalescent times specified by parameters in the MSC model with gene flow, such as species divergence times, population sizes, and rates of gene flow (Barton, 2006; Lohse and Frantz, 2014). As a result, there is important information concerning gene flow in the variance of site-pattern counts among loci, but this information is ignored by those methods. In other words, sites at the same locus share the genealogical history under the assumption of no within-locus recombination (see Zhu *et al.*, 2022 for an evaluation of the impact of this assumption on MSC-based analyses), and their differences reflect the stochastic fluctuation of the mutation process. Sites at different loci in addition may have different genealogical histories, reflecting the stochastic nature of the process of coalescent and introgression. When sites are pooled across loci, those two sources of variation are confounded, leading to loss of information (Shi and Yang, 2018; Zhu and Yang, 2021). As a consequence, certain forms of introgression, such as introgression between sister lineages, are unidentifiable by *D* or HyDe, while estimation of introgression rates between non-sister species suffers from larger variances (Jiao *et al.*, 2021).

Second, HyDe makes restrictive assumptions about gene flow. The underlying model is one of hybrid speciation with identical population sizes or equivalently the inflow model with symmetrical species divergence times and population sizes (fig. 2.7**a**, with $\tau_S = \tau_T$ and $\theta_S = \theta_T$) (Blischak *et al.*, 2018; Kubatko and Chifman, 2019). Our simulation suggests that HyDe can indeed infer gene flow/hybridization and produce reliable estimates of introgression probability under this model (fig. 2.9**d** & table S2.5; see also Blischak *et al.*, 2018; Flouri *et al.*, 2020). However, introgression in the wrong direction or violation of the symmetry assumptions may lead to loss of power and biased

or invalid estimates by HYDE (fig. 2.9**b**&**c**, table S2.5).

Third, the approaches taken by HYDE to accommodate multiple samples per species and heterozygote sites in diploid genomes may be problematic. When multiple samples are available in the species quartet, HYDE counts site patterns in all combinations of the quartet. Let the numbers of sequences for species O, P_1, H, P_2 be n_O, n_1, n_H, n_2. There are then $n_O \times n_1 \times n_H \times n_2$ combinations in which one sequence is sampled per species, and HYDE counts site patterns in all of them (Blischak *et al.*, 2018). This ignores the lack of independence among the quartets and exaggerates the sample size. At the same time, multiple samples from the same species are never compared with each other, which should provide important information about the population size for that species. In a likelihood method such as BPP, all sequences at the same locus, both from the same species and from different species, are related through a gene tree, and genealogical information at the locus is used.

Similarly heterozygote sites are not treated properly in HYDE. If the site pattern is AGRG, with R representing an A/G heterozygote, HYDE adds 0.5 each to the site patterns $ijjj$ (for AGGG) and $ijij$ (for AGAG) (Blischak *et al.*, 2018), in effect treating R as an unknown nucleotide that is either A or G whereas correctly it means a heterozygote (both A and G). The proportion of heterozygotes in each diploid genome should be informative about θ for that population, but such information is not used by HYDE. In BPP, heterozygote sites are resolved into their underlying nucleotides using an analytical integration algorithm (so that R means both A and G, say), with the uncertainty in the genotypic phase of multiple heterozygous sites in a diploid sequence accommodated by averaging over all possible heterozygote phase resolutions, weighting them according to their likelihoods based on the sequence alignment at the locus (Flouri *et al.*, 2018; Gronau *et al.*, 2011). Simulations suggest that this approach has nearly identical statistical performance to using fully phased haploid genomic sequences (Gronau *et al.*, 2011; Huang *et al.*, 2022b).

In this chapter we have focused on the heuristic method HYDE and the likelihood method BPP, as they have been used to analyse the *Tamias* data. By choosing parameter values to be representative of the *Tamias* data, our simulation has evaluated a tiny portion of the parameter space and does not constitute a systematic evaluation of the performance of HYDE. The strengths and weaknesses of heuristic and likelihood methods for inference under models of gene flow were discussed by Degnan

(2018) and Jiao *et al.* (2021), but a comprehensive comparative study has not yet been conducted. For estimation of the species phylogeny under the MSC without gene flow, (Zhu and Yang, 2021, fig. 3) demonstrated a dramatic information loss resulting from pooling sites across loci in the site-pattern based methods (also known as coalescent-aware concatenation methods), and from the failure to use information in coalescent times or gene-tree branch lengths in the two-step methods (which infer the gene trees and then treat them as data to infer the species tree). Both the site pattern-based and the two-step methods are used to infer gene flow and to estimate the introgression probability (e.g., HYDE and the *D*-statistic in the first category and SNAQ in the second) and similar information loss may be expected. A detailed analysis of the performance of heuristic methods in comparison with likelihood methods will be interesting. Currently the gap between the heuristic and likelihood methods appears to be a large one. Heuristic methods are orders-of-magnitude more efficient computationally and can be applied to much larger datasets, whereas likelihood methods have far better statistical properties, being able to identify and estimate all parameters in the model. There are great opportunities for improving both the statistical performance of heuristic methods and the computational efficiency of likelihood methods (including the mixing efficiency of MCMC algorithms).

2.4.3 Introgression in *T. quadrivittatus* chipmunks

The joint introgression model for the *T. quadrivittatus* group (fig. 2.4**b**) was constructed using a stepwise approach that iteratively adds introgression events to the binary species tree. We note several limitations with this approach. First the approach assumes the availability of a stable binary species tree, and may not be feasible if the species tree is large and highly uncertain, possibly influenced by introgression events (Leaché *et al.*, 2014). The *Tamias* dataset analysed here includes only six species, and the first stage of our procedure (i.e., the separate analysis) involved 16 possible introgression events, so that the computation was feasible. Second, the approach is not an exhaustive search in the space of introgression models and may miss certain introgression events. Note that introgression events not selected in the first stage of the procedure will not be incorporated in the final joint introgression model. In our analysis of the *Tamias* data, we considered introgressions between contemporary species, mostly based on phylogenetic analyses of the mitochondrial genome (Sarver

et al., 2017), and moved certain events to older ancestral branches when the estimated introgression time coincided with the species divergence time. We did not evaluate introgressions involving ancestral branches systematically. Furthermore, the criterion based on the Bayes factor used in our test is a stringent one, and the dataset of 1060 loci is relatively small. All those factors suggest that we cannot rule out the possibility that we may have missed some introgression events; in other words, our analysis may suffer from false-negative errors. In contrast, the three introgression events identified in our analysis (fig. 2.4**b**) appear to be robust and are unlikely to be false positives (figs. 2.5, table S2.2). We conclude that there is strong and robust evidence that gene flow has affected the nuclear genome in the *T. quadrivittatus* group of chipmunks.

Given the extensive mitochondrial introgression in the *Tamias* group (Sarver *et al.*, 2017, 2021; Sullivan *et al.*, 2014), introgression affecting the nuclear genome was expected, and the failure to detect any significant evidence for it in the HYDE analysis was surprising (Sarver *et al.*, 2021). Sarver *et al.* (2021) discussed the evidence for cytonuclear discordance in the pattern of introgression (Bonnet *et al.*, 2017; McElroy *et al.*, 2020; Sarver *et al.*, 2021), as well as possible roles of purifying selection affecting the coding genes or exons that make up the nuclear dataset being analysed. Our results suggest a simpler explanation, that gene flow in the *Tamias* group is of a wrong type or in the wrong direction, undetectable by HYDE.

Our analyses suggest that species involved in excessive mitochondrial introgression tend to be those involved in nuclear introgression as well. *T. dorsalis* was noted to be a universal recipient of mtDNA from other species (Sarver *et al.*, 2017; Sullivan *et al.*, 2014). Consistent with this, our separate analysis (table 2.2) identified three introgression events into *T. dorsalis* with $\varphi > 5\%$ as well as one event with *T. dorsalis* to be the donor species, even though some of those events become non-significant after introgression involving older ancestors was incorporated in the model. It will be interesting to use expanded datasets to examine whether this is due to a lack of power to detect gene flow or a genuine lack of gene flow.

It will be very useful to generate more genomic data, especially the noncoding parts of the nuclear genome, including more species from the genus, to provide more power for detecting gene flow and estimating introgression rates. It will also be interesting to examine whether the noncoding and coding

regions of the genome give consistent signals concerning species divergences and cross-species gene flow, and to examine how the effective rate of gene flow vary among chromosomes or across genomic regions. In a few genomic analyses, coding and noncoding parts of the genome were found to produce highly consistent results, with nearly proportional estimates of divergence times (τ) and population sizes (θ), and with very similar estimates of introgression rates (Shi and Yang, 2018; Thawornwattana *et al.*, 2018, 2022). One can also examine the posterior distribution of the gene trees to identify loci or genomic segments that are most likely to have been transferred across species boundaries, and to correlate with the functions of genes residing in or tightly linked to the segments.

2.5 Supplemental Information

Table S2.1: Posterior means and 95% HPD CIs (in parentheses) of introgression probabilities (φ) and introgression times (τ) in the stepwise construction of the MSci model, applied to datasets of the two halves

		First half			Second half		
	Model	φ	τ ($\times 10^{-3}$)	B_{10}	φ	τ ($\times 10^{-3}$)	B_{10}
1	QIRCD → U	0.537 (0.247, 0.801)	0.841 (0.702, 0.980)	∞	0.632 (0.332, 0.869)	0.906 (0.740, 1.045)	∞
2	QIRCD → U	0.615 (0.427, 0.798)	0.869 (0.773, 0.966)	∞	0.695 (0.501, 0.876)	0.918 (0.802, 1.037)	∞
	QI ↔ D	0.138 (0.094, 0.185)	0.349 (0.311, 0.391)	∞	0.069 (0.038, 0.102)	0.322 (0.282, 0.363)	∞
		0.020 (0.000, 0.047)		0.03	0.018 (0.000, 0.037)		0.03
3	QIRCD → U	0.601 (0.369, 0.813)	0.863 (0.759, 0.971)	∞	0.696 (0.480, 0.892)	0.915 (0.782, 1.040)	∞
	QI → D	0.133 (0.080, 0.191)	0.331 (0.288, 0.372)	53.24	0.085 (0.038, 0.136)	0.331 (0.279, 0.388)	25.06
	Q → D	0.008 (0.000, 0.021)	0.153 (0.058, 0.261)	0.00	0.008 (0.001, 0.019)	0.100 (0.040, 0.210)	0.00
4	QIRCD → U	0.588 (0.360, 0.797)	0.854 (0.746, 0.958)	∞	0.681 (0.470, 0.869)	0.909 (0.787, 1.032)	∞
	QI → D	0.132 (0.080, 0.188)	0.330 (0.286, 0.373)	∞	0.088 (0.035, 0.149)	0.334 (0.270, 0.396)	35.99
	I → D	0.007 (0.000, 0.018)	0.120 (0.048, 0.197)	0.00	0.012 (0.001, 0.025)	0.160 (0.090, 0.227)	0.01
5	QIRCD → U	0.582 (0.326, 0.818)	0.852 (0.738, 0.961)	∞	0.689 (0.473, 0.883)	0.905 (0.778, 1.026)	∞
	QI → D	0.126 (0.084, 0.170)	0.307 (0.263, 0.352)	∞	0.099 (0.063, 0.139)	0.336 (0.291, 0.382)	∞
	Q ↔ I	0.036 (0.013, 0.065)	0.099 (0.062, 0.137)	2.90	0.048 (0.023, 0.075)	0.088 (0.051, 0.118)	∞
		0.030 (0.008, 0.055)		0.39	0.014 (0.000, 0.028)		0.02
6	QIRCD → U	0.589 (0.338, 0.827)	0.850 (0.738, 0.966)	∞	0.686 (0.491, 0.870)	0.902 (0.782, 1.022)	∞
	QI → D	0.118 (0.074, 0.165)	0.295 (0.246, 0.345)	∞	0.097 (0.060, 0.136)	0.334 (0.284, 0.381)	∞
	Q → I	0.041 (0.014, 0.074)	0.100 (0.066, 0.140)	7.90	0.055 (0.026, 0.087)	0.091 (0.053, 0.132)	∞

Note.— Introgression events are added sequentially onto the species tree of figure 2.4**a** and those that do not meet our cutoffs ($B_{10} \geq 20$) are greyed out. $B_{10} = \infty$ occurs when there are no MCMC samples with $\varphi < \varepsilon = 1\%$. A bidirectional introgression event, e.g., between Q and I has two introgression probabilities, e.g., $\varphi_{Q \to I}$ (above) and $\varphi_{I \to Q}$ (below). The final joint introgression model has three unidirectional introgression events.

Table S2.2: Bayes factors (B_{10}) for the three introgression probabilities (φ) obtained from BPP analyses of the full data of 1060 loci under the joint MSci model of figure 2.4**b** and different beta priors, $\varphi \sim$beta(α, β)

		B_{10}		
ε & Prior	$\mathbb{P}(\text{ø})$	QIRCD $\rightarrow$ U	QI $\rightarrow$ D	Q $\rightarrow$ I
$\varepsilon = 1\%$				
beta(0.2, 0.2)	0.210	∞	∞	∞
beta(0.2, 1)	0.398	∞	∞	∞
beta(0.2, 5)	0.585	∞	∞	∞
beta(1, 0.2)	0.002	∞	∞	∞
beta(1, 1)	0.010	∞	∞	∞
beta(1, 5)	0.049	∞	∞	∞
beta(5, 0.2)	6.0×10^{-12}	∞	∞	∞
beta(5, 1)	1.0×10^{-10}	∞	∞	∞
beta(5, 5)	1.2×10^{-8}	∞	∞	∞
$\varepsilon = 0.1\%$				
beta(0.2, 0.2)	0.132	∞	∞	∞
beta(0.2, 1)	0.251	∞	∞	∞
beta(0.2, 5)	0.371	∞	∞	∞
beta(1, 0.2)	2.0×10^{-4}	∞	∞	∞
beta(1, 1)	0.001	∞	∞	∞
beta(1, 5)	0.005	∞	∞	∞
beta(5, 0.2)	6.0×10^{-17}	∞	∞	∞
beta(5, 1)	1.0×10^{-15}	∞	∞	∞
beta(5, 5)	1.3×10^{-13}	∞	∞	∞

Note.— Bayes factor B_{10} is calculated using eq. 2.8, where the null region ø for φ is the interval $(0, \varepsilon)$ with $\varepsilon = 1\%$ or 0.1%. $B_{10} = \infty$ occurs when $\varphi > \varepsilon$ in all MCMC samples.

Table S2.3: Posterior means and 95% HPD CIs (in parentheses) of parameters under the MSci model of figure 2.4**b** obtained from BPP analyses of three real datasets (the two halves and the full dataset) and a simulated dataset

Parameters	First half, 530 loci	Second half, 530 loci	Full data, 1060 loci	Simulation, 1060 loci
Population sizes (θ, $\times 10^{-3}$)				
θ_Q	1.119 (0.817, 1.455)	1.032 (0.733, 1.370)	1.059 (0.844, 1.287)	1.098 (0.867, 1.347)
θ_I	1.556 (0.996, 2.191)	2.335 (1.474, 3.387)	1.923 (1.433, 2.473)	2.108 (1.574, 2.701)
θ_R	0.330 (0.266, 0.399)	0.377 (0.311, 0.442)	0.344 (0.295, 0.396)	0.366 (0.313, 0.421)
θ_C	0.478 (0.400, 0.556)	0.474 (0.407, 0.547)	0.478 (0.427, 0.534)	0.491 (0.436, 0.542)
θ_D	3.092 (2.580, 3.633)	3.460 (2.920, 4.016)	3.314 (2.915, 3.705)	3.386 (3.001, 3.781)
θ_U	0.953 (0.843, 1.063)	0.912 (0.814, 1.011)	0.932 (0.857, 1.004)	0.917 (0.844, 0.991)
θ_S	0.792 (0.680, 0.904)	0.934 (0.809, 1.052)	0.866 (0.782, 0.948)	0.817 (0.734, 0.900)
$\theta_{QIRCDUS}$	11.04 (9.516, 12.56)	10.83 (9.357, 12.30)	11.01 (9.924, 12.09)	10.38 (9.368, 11.40)
θ_{QIRCDU}	0.687 (0.332, 1.075)	0.601 (0.274, 0.955)	0.656 (0.367, 0.971)	0.475 (0.229, 0.734)
θ_{QIRCD}	1.963 (0.248, 4.652)	1.595 (0.336, 3.153)	2.203 (0.392, 4.533)	1.340 (0.236, 2.809)
θ_{QIRC}	3.212 (0.296, 6.828)	1.503 (0.232, 3.505)	2.222 (0.295, 4.800)	2.141 (0.192, 5.173)
θ_{QIR}	2.923 (0.724, 5.067)	1.990 (0.755, 3.258)	2.518 (1.266, 3.890)	2.792 (1.523, 4.167)
θ_{QI}	0.727 (0.198, 1.503)	1.033 (0.203, 2.427)	0.773 (0.177, 1.714)	1.157 (0.217, 2.714)
θ_J	1.017 (0.777, 1.272)	1.242 (0.954, 1.545)	1.107 (0.921, 1.298)	1.147 (0.934, 1.372)
θ_K	0.686 (0.177, 1.467)	0.799 (0.177, 1.808)	0.626 (0.170, 1.282)	0.959 (0.185, 2.241)
θ_L	1.911 (0.224, 4.493)	1.280 (0.399, 2.326)	1.568 (0.345, 2.936)	1.381 (0.315, 2.781)
θ_M	0.412 (0.245, 0.569)	0.430 (0.282, 0.578)	0.407 (0.275, 0.529)	0.415 (0.291, 0.543)
θ_N	0.439 (0.310, 0.574)	0.476 (0.350, 0.600)	0.440 (0.342, 0.543)	0.384 (0.301, 0.473)
θ_O	0.291 (0.190, 0.390)	0.422 (0.282, 0.552)	0.325 (0.239, 0.416)	0.350 (0.259, 0.443)
Speciation/introgression times (τ, $\times 10^{-3}$)				
$\tau_{QIRCDUS}$	3.297 (2.830, 3.851)	3.588 (3.062, 4.075)	3.423 (3.061, 3.783)	3.415 (3.066, 3.768)
τ_{QIRCDU}	1.872 (1.299, 2.456)	2.270 (1.703, 2.849)	2.029 (1.569, 2.489)	2.011 (1.673, 2.338)
τ_{QIRCD}	0.749 (0.634, 0.855)	0.750 (0.654, 0.837)	0.731 (0.642, 0.815)	0.753 (0.693, 0.815)
τ_{QIRC}	0.584 (0.469, 0.699)	0.673 (0.573, 0.765)	0.628 (0.542, 0.707)	0.697 (0.615, 0.766)
τ_{QIR}	0.363 (0.283, 0.452)	0.437 (0.360, 0.517)	0.389 (0.322, 0.452)	0.379 (0.312, 0.445)
τ_{QI}	0.267 (0.222, 0.312)	0.321 (0.272, 0.367)	0.290 (0.253, 0.327)	0.274 (0.238, 0.309)
$\tau_J = \tau_K = \tau_{QIRCD \to U}$	0.850 (0.738, 0.966)	0.902 (0.782, 1.022)	0.871 (0.778, 0.961)	0.832 (0.747, 0.917)
$\tau_L = \tau_M = \tau_{QI \to D}$	0.295 (0.246, 0.345)	0.334 (0.284, 0.381)	0.307 (0.268, 0.350)	0.298 (0.258, 0.336)
$\tau_N = \tau_O = \tau_{Q \to I}$	0.100 (0.066, 0.140)	0.091 (0.053, 0.132)	0.102 (0.074, 0.130)	0.094 (0.069, 0.118)
Introgression probabilities (φ)				
$\varphi_{QIRCD \to U}$	0.589 (0.338, 0.827) (∞)	0.686 (0.491, 0.870) (∞)	0.625 (0.442, 0.794) (∞)	0.587 (0.440, 0.733) (∞)
$\varphi_{QI \to D}$	0.118 (0.074, 0.165) (∞)	0.097 (0.060, 0.136) (∞)	0.106 (0.074, 0.139) (∞)	0.107 (0.077, 0.140) (∞)
$\varphi_{Q \to I}$	0.041 (0.014, 0.074) (8)	0.055 (0.026, 0.087) (∞)	0.050 (0.028, 0.074) (∞)	0.048 (0.028, 0.069) (∞)

Note.— Bayes factor B_{10} is given in parentheses, calculated using eq. 2.8: ∞ means that all sampled values of φ are $> \varepsilon = 1\%$.

Table S2.4: Power of BPP, HYDE and *D*-statistic tests of gene flow between sister species and average estimates of introgression probability in 100 simulated replicate datasets (each of 8000 loci) under the model of figure 2.8**a**

	Power			Proportion of
Methods	($\alpha = 1\%$)	($\alpha = 5\%$)	$\hat{\varphi}\pm$SD	invalid estimates
HYDE				
R→Ua	3%	8%	0.005 ± 0.003	48%
R→Ub	3%	5%	0.004 ± 0.004	51%
D-statistic				
R↔Ua	0%	1%	–	NA
R↔Ub	0%	2%	–	NA
BPP				
R→U	100%	100%	0.623 ± 0.066	0%

Note.— Bayesian test by BPP is considered significant at the 5% (or 1%) level if $B_{10} \geq 20$ (or 100). In the HYDE test, Ua and Ub were regarded as the 'hybrid' lineage to detect gene flow R→Ua and R→Ub, respectively, in figure 2.8**a**. In some datasets, the HYDE estimate of φ was outside the range (0, 1), and only the valid estimates were used to calculate the means.

Table S2.5: Power of BPP and HYDE tests of gene flow and average estimates of introgression probability in 100 simulated replicates under the three models of figure 2.8**b**-**d**

	BPP			HYDE			
# loci	Power ($\alpha = 1\%$)	Power ($\alpha = 5\%$)	$\hat{\varphi}\pm$SD	Power ($\alpha = 1\%$)	Power ($\alpha = 5\%$)	$\hat{\varphi}\pm$SD	Proportion of invalid estimates
Outflow asym (fig. 2.8**b**)							
500	39%	56%	0.096 ± 0.026	1%	4%	0.155 ± 0.111	44%
2000	100%	100%	0.104 ± 0.025	3%	9%	0.107 ± 0.057	33%
8000	100%	100%	0.105 ± 0.013	10%	24%	0.076 ± 0.042	20%
Inflow asym (fig. 2.8**c**)							
500	72%	84%	0.118 ± 0.030	23%	41%	0.331 ± 0.110	10%
2000	100%	100%	0.106 ± 0.014	87%	95%	0.321 ± 0.068	0%
8000	100%	100%	0.107 ± 0.009	100%	100%	0.325 ± 0.037	0%
inflow sym (fig. 2.8**b**, HYDE model)							
500	15%	27%	0.115 ± 0.037	2%	10%	0.124 ± 0.071	19%
2000	90%	95%	0.110 ± 0.022	14%	27%	0.101 ± 0.047	2%
8000	100%	100%	0.108 ± 0.010	83%	90%	0.108 ± 0.025	0%

Note.— The true introgression probability is $\varphi = 0.106$ (fig. 2.8**b**-**d**). See legend to table S2.4.

Chapter 3

Inference of Cross-Species Gene Flow Using Genomic Data Depends on the Methods: Case Study of Gene Flow in *Drosophila*

A recent phylogenomic analysis of protein-coding genes from *Drosophila* revealed widespread introgression across a phylogeny of 149 species (Suvorov *et al.*, 2022). The data were split into nine well-supported clades to detect gene flow within each. Several tests based on rooted triplets (or unrooted quartets) were employed, including two newly developed approaches: the discordant count test (DCT) and branch length test (BLT) (Suvorov *et al.*, 2022). Applied to species triplets, DCT appears to be equivalent to SNAQ (Solis-Lemus and Ane, 2016; Solis-Lemus *et al.*, 2017), while BLT is similar to QUIBL as both use estimated branch lengths in triplet gene trees. Another method used by Suvorov *et al.* (2022) is PHYLONET (Wen *et al.*, 2018), which takes inferred gene-tree topologies as input data and ignores information in coalescent times. Those methods cannot identify gene flow between sister lineages and cannot identify the direction of gene flow. As gene flow involving ancestral species may show up in many triplet tests, a heuristic metric called f-branch was used to move introgression events to ancestral branches in the given species tree (Malinsky *et al.*, 2018). The approach does not consider species divergence times or introgression times, and may assign gene flow to donor and recipient populations that were not contemporary. Such limitations of the analytical methods used by Suvorov *et al.* (2022) suggest a need for reanalysis of the data using likelihood methods such as BPP. In a recent analysis of exonic data from six Rocky Mountain chipmunk species in the *Tamias* group, the summary method HYDE failed to detect any signal of gene flow affecting the nuclear genome, in contrast to the mitochondrial genome, which is well-known to be involved in rampant gene flow in the group, prompting discussions of cytonuclear discordance (Sarver *et al.*, 2021). However, a reanalysis of the same data using BPP detected robust evidence for multiple ancient introgression events affecting the nuclear genome, including one between sister

species (Ji *et al.*, 2023), suggesting no evidence for cytonuclear discordance. Thus analyses of the same data using summary (Sarver *et al.*, 2021) and Bayesian (Ji *et al.*, 2023) methods produced opposing biological conclusions. It is unclear whether the conclusions of Suvorov *et al.* (2022) are similarly affected by the use of summary methods.

Here we apply the MSC-I and MSC-M models implemented in BPP (Flouri *et al.*, 2020, 2023) to reanalyse a subset of the *Drosophila* data of Suvorov *et al.* (2022). We used data from clade 2, which showed the strongest signal of introgression in the analysis of Suvorov *et al.* (2022, Table 1). Consistent with Suvorov *et al.* (2022), we detected strong evidence for gene flow, but the details differ. The strongest signature of introgression in our analysis is between two sister lineages, not detected by Suvorov *et al.* (2022), while several gene-flow scenarios inferred by Suvorov *et al.* (2022) are rejected in our test. To understand the differences in the results from the two studies, we conduct computer simulations to evaluate the statistical properties of BPP and the summary methods used by Suvorov *et al.* (2022), including HyDe, QuIBL, DCT, BLT, and SNaQ. Our results suggest that the different results may be explained by the lack of power of the summary methods used. Our study highlights the need and importance of using powerful statistical methods to infer gene flow using genomic datasets.

3.1 Materials and Methods

3.1.1 The *Drosophila* dataset

Suvorov *et al.* (2022) generated and compiled sequence alignments for 2794 single-copy protein-coding genes (BUSCO, for Benchmarking Universal Single-Copy Orthologs) from 155 *Drosophila* species and constructed a species phylogeny. Data for nine well-established clades were then used to infer interspecific gene flow. Here we used data for clade 2 in the species tree, comprised of 11 species: *D. affinis, D. athabasca, D. azteca, D. lowei, D. miranda, D. persimilis, D. pseudoobscura, D. bifasciata, D. obscura, D. guanche* and *D. subobscura* (Fig. 3.1a). Seventeen loci had <2 species and were removed, leaving 2777 loci. The 2777 loci were split into two random halves, with 1389 and 1388 loci, respectively, and analysed separately.

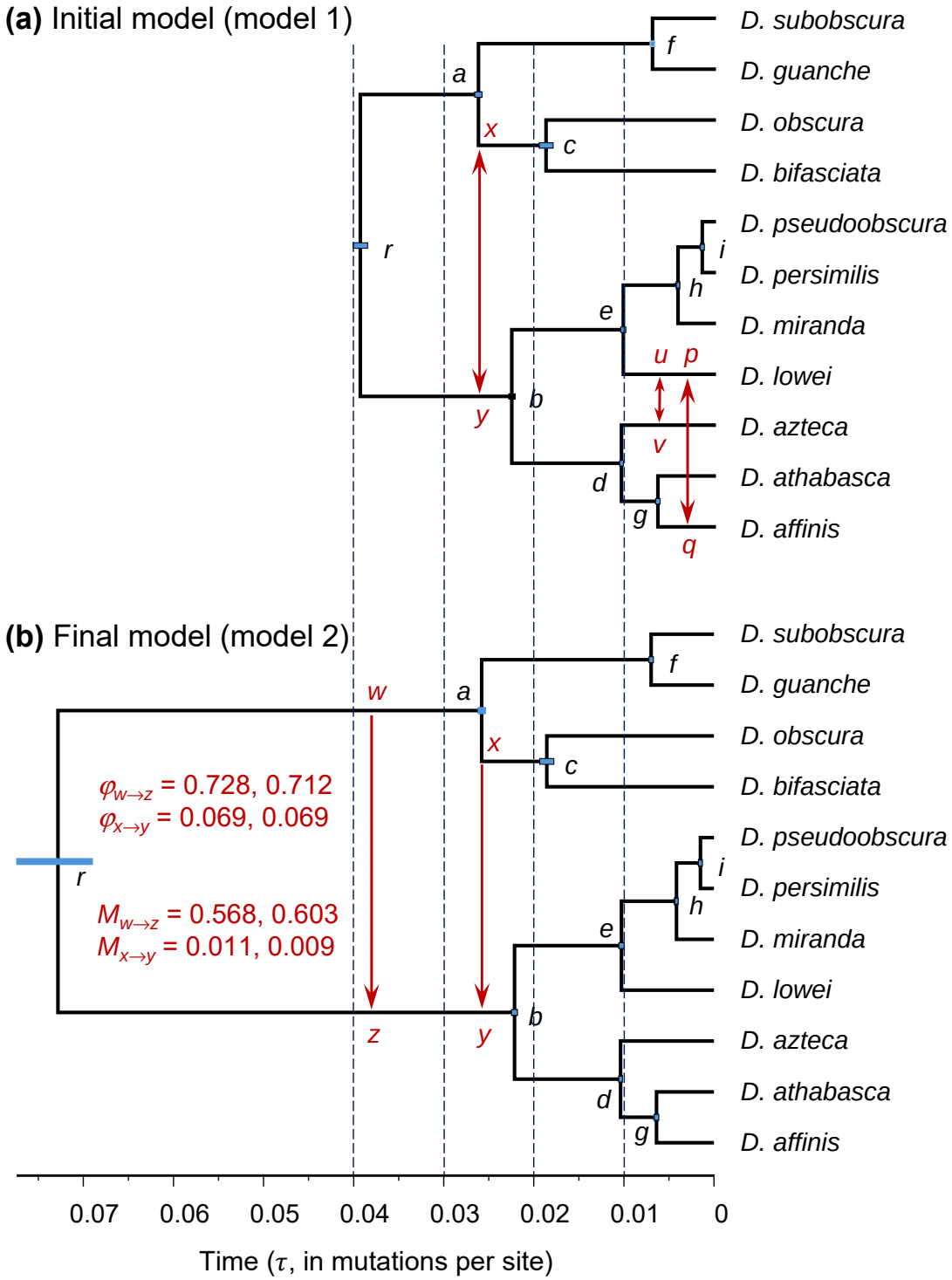

Figure 3.1: (**a**) Species phylogeny for 11 *Drosophila* species in clade 2 of Suvorov *et al.* (2022) showing potential gene-flow events in our initial model. Arrows represent potential gene-flow events, based on analyses of species triplets by Suvorov *et al.* (2022) (Table S3.1) and on BPP estimates of divergence times. (**b**) Final model of gene flow from our analysis, with two gene-flow events from branches *ra* to *rb* (that is, $w \to z$), and from *ac* to *rb* ($x \to y$). Estimates of the introgression probability (φ) in the MSC-I model and migration rate (M) in the MSC-M model are from the two data halves. Branch lengths are proportional to posterior means of species divergence times and introgression times (τ, measured in mutations per site) with node bars representing the 95% HPD CIs, from BPP analyses of the first half of the data under the MSC-I model. Estimates for the second half are very similar. Estimates of all parameters under both the MSC-I and MSC-M models for the two data halves are in Table S3.4.

3.1.2 Inferring the *Drosophila* species phylogeny

We inferred the species tree under the MSC model with no gene flow using BPP (Rannala and Yang, 2017; Yang and Rannala, 2014). This is the A01 analysis of Yang (2015). The two data halves were analysed separately. There are two types of parameters in the MSC model: species divergence times (τ) and population sizes (θ), both measured in the expected number of mutations per site. We assigned the gamma prior to the age of the species-tree root, $\tau_R \sim \text{G}(2, 50)$, with mean $2/50 = 0.04$. Given the age of the root, the other divergence times had the uniform-Dirichlet prior distribution (Yang and Rannala, 2010, eq. 2). A gamma prior is assigned to population size parameters on the species tree, $\theta \sim \text{G}(2, 200)$, with mean 0.01. The JC mutation model (Jukes and Cantor, 1969) was used in calculation of the likelihood for the sequence alignment at each locus. We expect JC to be sufficient for correcting for multiple hits at the same site because sequences from closely related species are highly similar (Flouri *et al.*, 2022; Shi and Yang, 2018) (see below for further tests). We used a burn-in of 40,000 MCMC iterations, and then took 2×10^5 samples, sampling every 2 iterations. Each analysis was repeated four times, with convergence of the MCMC confirmed by consistency across runs.

3.1.3 Constructing a model of gene flow for the *Drosophila* data

Species tree inference using BPP produced a well-supported species phylogeny, which had the same topology as inferred by Suvorov *et al.* (2022) (Figs. 3.1a & S3.1). The species phylogeny appeared to be unaffected by gene flow. We thus added candidate gene-flow events onto this binary species tree, using a procedure similar to that followed by Ji *et al.* (2023) in their analysis of a chipmunk genomic dataset. We assessed the gene-flow scenarios proposed by Suvorov *et al.* (2022, Fig. 3) by integrating their DCT/BLT analyses of many species triplets (Table S3.1), with reference to estimated species divergence times from BPP. The triplet methods of Suvorov *et al.* (2022) are unable to identify the direction of gene flow (e.g., Pang and Zhang, 2024; Thawornwattana *et al.*, 2023a). Thus we assumed bidirectional gene flow in our initial model, with the expectation that if the gene-flow event in a particular direction is nonexistent, the estimated rate of gene flow will be close to zero and the Bayesian test will reject gene flow (Thawornwattana *et al.*, 2023a). The resulting initial model of

gene flow is shown in Figure 3.1a.

We then applied the Bayesian test of gene flow (Ji *et al.*, 2023) to determine the significance of the gene-flow events in the model. While Ji *et al.* (2023) sequentially added introgression events onto the species tree, starting from the most significant introgression events, we fitted the full model with all gene-flow events, and used the Bayes factor to remove events that are not strongly supported by the data. The Bayes factor B_{10}, in support of the alternative model of gene flow (H_1) against the null model of no gene flow (H_0), was calculated via the Savage-Dickey density ratio using an MCMC sample under the H_1 model (Ji *et al.*, 2023). Gene flow was accommodated using either the MSC-I or MSC-M models. Under MSC-I, the strength of gene flow is measured by the introgression probability, φ_{XY}, which is the proportion of immigrants in the recipient population Y from X. We defined a 'null interval' for the introgression probability, $\varphi < \varepsilon$, which is a small interval in the parameter space of H_1 that represents H_0. Then B_{10} is approximated by

$$B_{10,\varepsilon} = \frac{\mathbb{P}(\varphi < \varepsilon)}{\mathbb{P}(\varphi < \varepsilon | X)}, \tag{3.1}$$

where $\mathbb{P}(\varphi < \varepsilon)$ and $\mathbb{P}(\varphi < \varepsilon|X)$ are the prior and posterior probabilities for $\varphi < \varepsilon$, respectively. When $\varepsilon \to 0$, $B_{10,\varepsilon} \to B_{10}$ (Ji *et al.*, 2023). We used $\varepsilon = 0.01$ and confirmed that use of $\varepsilon = 0.001$ gave similar results. We used a cut-off of 100. Thus $B_{10} > 100$ means strong support for H_1 and rejection of H_0, which is similar to significance at the 1% level in hypothesis testing. $B_{10} < 0.01$ means strong support for H_0 and rejection of H_1. This does not have an equivalence in hypothesis testing as hypothesis testing can never reject H_1 with great force. See Ji *et al.* (2023) for detailed discussions.

Under the MSC-M model, the population migration rate, $M_{XY} = m_{XY}N_Y$, is defined as the expected number of migrants from the donor species X to the recipient species Y per generation, where m_{XY} is the proportion of migrants in Y from X and N_Y is the (effective) population size of species Y. Bayes factor B_{10} in support of $H_1 : M > 0$ against the null $H_0 : M = 0$ was calculated by defining a null interval $M < \varepsilon$, with $\varepsilon = 0.01$ or 0.001.

Thus calculation of B_{10} using eq. 3.1 requires running the MCMC under the model of gene flow

(H_1). Under the MSC-I, the introgression probability was assigned the prior $\varphi \sim \text{beta}(1,1)$ or $\mathbb{U}(0,1)$. We used the option `theta-model = linked-msci` in BPP, which assumes the same population-size parameter θ for a branch before and after an introgression event (Ji *et al.*, 2023, Fig. 3b). Under the MSC-M, the migration rate was assigned the gamma prior $M \sim \text{G}(2,10)$, with mean 0.2. We used a burn-in of 10^5 iterations, after which we took 5×10^5 samples, sampling every 2 iterations. Each analysis was conducted four times to confirm convergence, indicated by the difference in the posterior probability for the *maximum a posteriori* (MAP) tree between runs being less than 0.3 (Thawornwattana *et al.*, 2022). Runs that did not converge were discarded before the MCMC samples from multiple runs were combined to produce posterior summaries. Each MSC-I run took ~90hrs using two threads, while each MSC-M run took ~120hrs using four threads.

Gene-flow events that passed the Bayesian test (with $B_{10} > 100$) are retained in the final model, which is then used to estimate population parameters, including the rates of gene flow (φ or M), species split times, and population sizes for extant and extinct species on the species tree.

3.1.4 Assessing the impact of taxon sampling

The evidence for gene flow involving *D. lowei* (see Fig. 3.1a) appeared to depend on the choice of the outgroup species and on other species included in the dataset. We thus constructed three triplet datasets and three quintet datasets, to assess the impact of taxon sampling. We focussed on gene flow between *D. lowei* and *D. affinis*, for which the evidence is significant in 2 out of 3 triplets in the analysis of Suvorov *et al.* (2022, Table 1).

For the triplet datasets, the species tree was ((*X*, *D. lowei*), *D. affinis*), where *X* was *D. pseudoobscura*, *D. persimilis*, or *D. miranda* (Fig. 3.1a). The data were also analysed using summary-based tests (DCT, BLT and QuIBL), with *D. guanche* used as the outgroup. For the quintet datasets, we included two outgroup species: *D. guanche* and *D. obscura*, so that the species tree was (((*X*, *D. lowei*), *D. affinis*), (*D. obscura*, *D. guanche*)), where *X* again was one of *D. pseudoobscura*, *D. persimilis*, or *D. miranda* (Fig. 3.1a). We applied the Bayesian test to assess the evidence for gene flow between *D. lowei* and *D. affinis*.

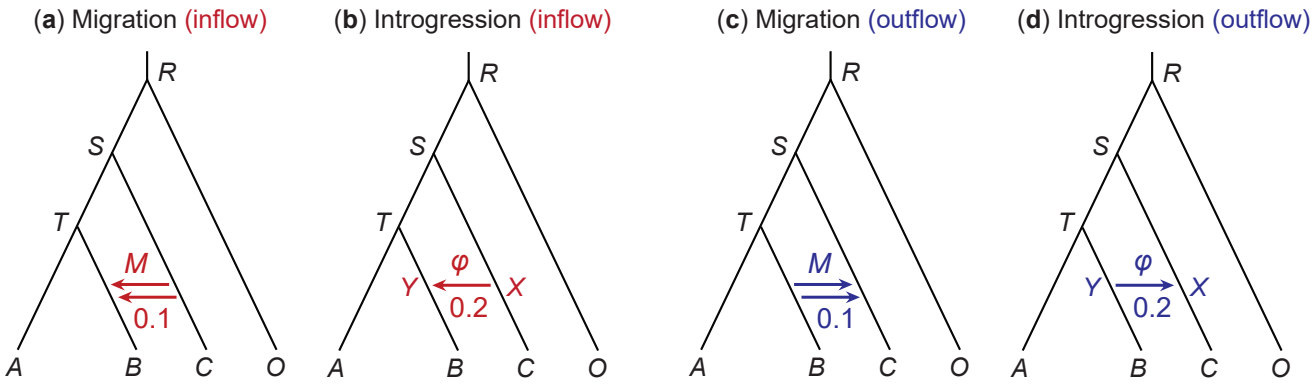

Figure 3.2: Migration (MSC-M) and introgression (MSC-I) models used to simulate and analyse multilocus sequence data. In the inflow models (**a** & **b**), gene flow is from $C \to B$, whereas in the outflow models (**c** & **d**), it is from $B \to C$. In the MSC-M model (**a** & **c**), migration occurs at the rate of $M = 0.1$ migrants per generation, whereas in the MSC-I model (**b** & **d**), the introgression probability is $\varphi = 0.2$. Species divergence times are $\tau_R = 3\theta$, $\tau_S = 2\theta$ and $\tau_T = \theta$. The introgression time under MSC-I is $\tau_X = \tau_Y = \theta/2$. Two values are used for the population size parameter: $\theta = 0.0025$ and 0.01. Each simulated dataset is analysed using BPP under both the MSC-M and MSC-I models, generating eight simulation-analysis combinations.

3.1.5 Simulating data to evaluate Bayesian and summary methods for inferring gene flow

As our re-analysis of the *Drosophila* data (for clade 2) produced different results from those of Suvorov *et al.* (2022), we simulated data under the MSC model with gene flow to examine the accuracy of BPP estimation of parameters (Flouri *et al.*, 2020, 2023) and the power of Bayesian test of gene flow (Ji *et al.*, 2023), in comparison with the summary methods used by Suvorov *et al.* (2022).

We conducted two sets of simulations. In the first set, we simulated two datasets using parameter estimates obtained from the *Drosophila* data with the *D. insularis* outgroup under our final MSC-I and MSC-M models with the $w \to z$ and $x \to y$ gene-flow events, with parameter values given in Table S3.2 (first half) and Table S3.3 (first half). Each dataset consisted of 1388 loci, as in the original data halves. The `simulate` option in BPP (Flouri *et al.*, 2018) was used to generate data under the JC mutation model (Jukes and Cantor, 1969), which were then analysed using BPP under the same model.

In the second set of simulations, we used four artificial MSC-M and MSC-I models for four species (A, B, C, and outgroup O) of Figure 3.2, with gene flow between nonsister lineages to examine the performance of Bayesian test of gene flow and Bayesian estimation of the rate of gene flow, in comparison with summary methods. The four models of gene flow in Figure 3.2 were used to

simulate gene trees for the loci, which were then used to 'evolve' sequences under JC, resulting in a sequence alignment at each locus. The species divergence times were $\tau_R = 3\theta$, $\tau_S = 2\theta$ and $\tau_T = \theta$, with $\theta = 0.0025$ and 0.01. The migration rate was $M = 0.1$ under the MSC-M model (Fig. 3.2a & c). Under the MSC-I model, the introgression time was $\tau_X = \tau_Y = 0.5\theta$, and the introgression probability was $\varphi = 0.2$ (Fig. 3.2b & d).

We examined the effects of the number of loci ($L = 250, 1000, 4000$), the number of sequences per species per locus ($S = 2, 8$), the sequence length ($n = 250, 1000$), and mutation rate ($\theta = 0.0025, 0.01$). As the divergence times (τs) are proportional to θ in our experiment design, the two values of θ mimic genomic regions with different mutation rates (such as coding versus noncoding regions of the genome). We did not run BPP over the large datasets with $L = 4000$ loci and $S = 8$ sequences per species per locus, as those runs were expensive and BPP already achieved 100% power and highly precise parameter estimates in much smaller datasets. One hundred replicates were generated for each parameter setting, with a total of 2000 ($= 3 \times 2 \times 2 \times 2 \times 100 - 400$) datasets generated for each of the four models of Figure 3.2.

Each replicate dataset (simulated under either MSC-I or MSC-M) was analysed using BPP under both the MSC-M and MSC-I models, resulting in eight simulation-analysis settings. When the data were analysed, the correct source and donor populations were assumed in the MSC model with gene flow. Gamma priors were assigned to the population size parameters (θ) and the age of the species-tree root (τ_R). We used the shape parameter $\alpha = 2$ and adjusted the rate parameter (β) so that the prior means are equal to the true values. For example, for data simulated using $\theta = 0.0025$ in the M-M and M-I settings, we used the priors $\theta \sim \text{G}(2, 800)$ and $\tau_0 \sim \text{G}(2, 266)$, whilst for data simulated using $\theta = 0.01$, we used $\theta \sim \text{G}(2, 200)$ and $\tau_0 \sim \text{G}(2, 66)$. Note that while the same θ was used for all populations when data were simulated, each branch on the species tree had its own θ when the data were analysed. Under the MSC-I model, we used the `thetamodel = linked-msci` option so that the same population size parameter is assumed for a branch before and after introgression. Additionally, we used the priors $\varphi \sim \text{beta}(1, 1)$ under MSC-I and $M \sim \text{G}(2, 20)$ under MSC-M. A burn-in of 40,000 iterations was used, after which we took 10^5 samples sampling every 2 iterations.

We evaluated both the power of Bayesian test of gene flow (using the Bayes factor calculated

via the Savage-Dickey density ratio; see description above) and Bayesian estimation of parameters, including the rate of gene flow (φ in MSC-I and M in MSC-M). Performance in parameter estimation was measured using the width of the 95% highest probability density (HPD) credible interval (CI).

The simulated quartet data were also analysed using several summary methods, including those used by Suvorov *et al.* (2022). We assessed both the power to detect introgression and the bias and precision in estimation of the introgression probability. Methods used for testing introgression included HyDe (Blischak *et al.*, 2018), QuIBL (Edelman *et al.*, 2019), DCT (Suvorov *et al.*, 2022), and BLT (Suvorov *et al.*, 2022). Note that those methods are uninformative about the mode of gene flow (whether it occurs in a pulse or over an extended time period), and about the direction of gene flow, whilst BPP assumes a fully specified parametric model. Methods for estimating the introgression probability included HyDe, QuIBL, DCT, and SNaQ (Solis-Lemus and Ane, 2016). Those methods generate only point estimates of φ, while BPP provides in addition a measure of uncertainty in the posterior CIs.

HyDe was implemented using the python script `run_hyde.py` from Blischak *et al.* (2018) (`https://github.com/pblischak/HyDe`), which uses a concatenated alignment to count site patterns across all loci. DCT/BLT was implemented using `blt_dct_test.r` from Suvorov *et al.* (2022) (`https://github.com/SchriderLab/Drosophila_phylogeny`). QuIBL was run using `QuIBL.py` from Edelman *et al.* (2019) (`https://github.com/miriammiyagi/QuIBL`). For SNaQ, we used the PhyloNetworks package (Solis-Lemus and Ane, 2016) to estimate the introgression probability.

QuIBL, DCT, BLT and SNaQ were applied using gene trees for the individual loci reconstructed by RaxML with default settings (Stamatakis, 2014). Like SNaQ, DCT estimates the introgression probability using inferred gene tree topologies (Suvorov *et al.*, 2022):

$$\hat{\varphi} = \frac{c_{\text{dis2}} - c_{\text{dis1}}}{c_{\text{con}} + c_{\text{dis1}} + c_{\text{dis2}}},$$

where $c_{\text{con}}, c_{\text{dis1}}, c_{\text{dis2}}$ are the counts of concordant and discordant gene trees, with $\hat{\varphi} = 0$ if $c_{\text{dis1}} > c_{\text{dis2}}$.

QuIBL, DCT and BLT do not allow multiple sequences per species per locus. Thus input gene

trees were constructed using a single sequence chosen at random from among the two or eight sequences simulated for each species. Results were similar when different sequences were sampled.

Table 3.1: Posterior means and 95% HPD CIs (in parentheses) of introgression probabilities (φ), introgression times (τ) and Bayes factors in support of gene flow (B_{10}) in the BPP analysis of the *Drosophila* data under the MSC-I models of Figure 3.1

	First half (1389 loci)			Second half (1388 loci)		
Introgression	$\hat{\varphi}$	$\hat{\tau}$	B_{10}	$\hat{\varphi}$	$\hat{\tau}$	B_{10}
Model 1a: *D. lowei* ↔ *D. affinis* introgression first (Fig. 3.1a)						
rb → *ac* (or $y \to x$)	0.0014 (0.0000, 0.0041)	0.0261 (0.0257, 0.0265)	0.01	0.0011 (0.0000, 0.0032)	0.0260 (0.0256, 0.0265)	0.01
ac → *rb* (or $x \to y$)	0.0980 (0.0787, 0.1173)		∞	0.0917 (0.0756, 0.1081)		∞
D. lowei → *D. azteca*	0.0027 (0.0003, 0.0058)	0.0032 (0.0001, 0.0057)	0.01	0.0009 (0.0000, 0.0026)	0.0025 (0.0000, 0.0053)	0.01
D. azteca → *D. lowei*	0.0008 (0.0000, 0.0024)		0.01	0.0010 (0.0000, 0.0030)		0.01
D. lowei → *D. affinis*	0.0026 (0.0000, 0.0062)	0.0050 (0.0025, 0.0066)	0.01	0.0011 (0.0000, 0.0035)	0.0047 (0.0020, 0.0064)	0.01
D. affinis → *D. lowei*	0.0008 (0.0000, 0.0024)		0.01	0.0016 (0.0000, 0.0040)		0.01
Model 1b: *D. lowei* ↔ *D. azteca* introgression first (Fig. 3.1a)						
rb → *ac* (or $y \to x$)	0.0014 (0.0000, 0.0041)	0.0261 (0.0257, 0.0265)	0.01	0.0011 (0.0000, 0.0032)	0.0260 (0.0255, 0.0264)	0.01
ac → *rb* (or $x \to y$)	0.0980 (0.0788, 0.1173)		∞	0.0920 (0.0758, 0.1085)		∞
D. lowei → *D. azteca*	0.0026 (0.0001, 0.0058)	0.0063 (0.0032, 0.0102)	0.01	0.0010 (0.0000, 0.0029)	0.0075 (0.0034, 0.0103)	0.01
D. azteca → *D. lowei*	0.0008 (0.0000, 0.0025)		0.01	0.0014 (0.0000, 0.0039)		0.01
D. lowei → *D. affinis*	0.0020 (0.0000, 0.0051)	0.0033 (0.0001, 0.0061)	0.01	0.0011 (0.0000, 0.0033)	0.0034 (0.0028, 0.0062)	0.01
D. affinis → *D. lowei*	0.0008 (0.0000, 0.0024)		0.01	0.0013 (0.0000, 0.0036)		0.01
Model 2: final model with unidirectional introgression from $w \to z$ and $x \to y$ (Fig. 3.1b)						
ra → *rb* (or $w \to z$)	0.7275 (0.6893, 0.7690)	0.0381 (0.0375, 0.0387)	∞	0.7124 (0.6806, 0.7432)	0.0388 (0.0382, 0.0394)	∞
ac → *rb* (or $x \to y$)	0.0688 (0.0546, 0.0832)	0.0257 (0.0253, 0.0261)	∞	0.0690 (0.0561, 0.0822)	0.0257 (0.0253, 0.0261)	∞

Note.– Initial models 1a & 1b differ in the time order of two bidirectional introgression events: *D. lowei* ↔ *D. azteca* versus *D. lowei* ↔ *D. affinis* (Fig. 3.1a). As the time of the *ac* → *rb* introgression ($\tau_{x\to y}$) was very close to the species divergence time τ_a (Fig. 3.1a), the introgression event was moved to the parental branch in ($w \to z$, Fig. 3.1b), but there was support in the data for the $x \to y$ introgression, so that both events were included in the final model 2. Bayes factor for testing introgression (B_{10}) was calculated using the Savage-Dickey density ratio with $\varepsilon = 0.01$ (Ji *et al.*, 2023). $B_{10} = \infty$ occurs when there are no MCMC samples with $\varphi < \varepsilon = 0.01$, whereas $B_{10} = 0.01$ occurs when all MCMC samples have $\varphi < \varepsilon$.

3.2 Results

3.2.1 Inference of species tree and construction of an initial model of gene flow for the *Drosophila* data

Protein-coding genes from the 11 species in clade 2 of the *Drosophila* phylogeny of Suvorov *et al.* (2022) (Fig. 3.1a) were separated into two random subsets, with 1389 and 1388 loci, respectively. They were analysed separately using BPP to estimate the species tree under the MSC model with no gene flow (Flouri *et al.*, 2018; Yang, 2015). Analysis of the two data halves allowed us to assess the robustness of our results to the sampling of loci and also reduced the computational load. All runs

Table 3.2: Posterior means and 95% HPD CIs (in parentheses) of migration rates (M) and Bayes factors (B_{10}) in the BPP analysis of the *Drosophila* data under the MSC-M model of Figure 3.1

	First half (1389 loci)		Second half (1388 loci)	
Migration	$\hat{M}$	B_{10}	$\hat{M}$	B_{10}
Model 1 (Fig. 3.1a)				
$rb \to ac$ (or $y \to x$)	0.0220 (0.0028, 0.0449)	0.01	0.0075 (0.0002, 0.0181)	0.01
$ac \to rb$ (or $x \to y$)	0.3065 (0.2255, 0.3940)	∞	0.3375 (0.2588, 0.4212)	∞
D. lowei $\to$ *D. azteca*	0.0108 (0.0014, 0.0238)	0.00	0.0084 (0.0004, 0.0215)	0.01
D. azteca $\to$ *D. lowei*	0.0142 (0.0018, 0.0293)	0.01	0.0143 (0.0015, 0.0333)	0.02
D. lowei $\to$ *D. affinis*	0.0134 (0.0011, 0.0306)	0.00	0.0171 (0.0020, 0.0384)	0.00
D. affinis $\to$ *D. lowei*	0.0148 (0.0012, 0.0316)	0.02	0.0181 (0.0032, 0.0378)	0.01
Model 2: final model with unidirectional migration from $w \to z$ and $x \to y$ (Fig. 3.1b)				
$ra \to rb$ (or $w \to z$)	0.5677 (0.5183, 0.6151)	∞	0.6031 (0.5546, 0.6523)	∞
$ac \to rb$ (or $x \to y$)	0.0111 (0.0003, 0.0253)	0.01	0.0090 (0.0002, 0.0204)	0.01

Note.– Model 1 assumes three bidirectional migration events or three pairs of migration rates (Fig. 3.1a). All of them were rejected except $M_{x\to y}$ based on the Bayes factor (B_{10}). In the final model 2, we added the $w \to z$ migration, similarly to analysis under the MSC-I model (Table 3.1).

across the two halves produced the same species tree topology as inferred by Suvorov *et al.* (2022). We thus concluded that the species phylogeny was well established. The BPP analysis also produced Bayesian estimates of parameters including species divergence times (τ). This information was used, in conjunction with the introgression events inferred by Suvorov *et al.* (2022) in their analyses of triplet and quartet data, to construct an initial model of gene flow for clade 2.

Suvorov *et al.* (2022, Fig. 3) inferred three introgression events for clade 2 (Fig. S3.1). These were, with nodes and branches labelled as in Figure 3.1a: (i) between x and y, (ii) between branches *be* and *af*, and (iii) between lineages *bd* and *D. lowei*. Event ii had only weak support, with significant evidence for gene flow in only two out of 40 feasible triplets (Fig. S3.1). This was thus discarded in our initial model. Event iii involved branch *bd* and the *D. lowei* lineage (Fig. 3.1a), inferred by Suvorov *et al.* (2022) using the f-branch approach (Malinsky *et al.*, 2018). These two lineages did not appear to overlap in time according to BPP estimates of species divergence times. While such a scenario could be interpreted as introgression involving an extinct or unsampled "ghost" lineage (e.g., Yang and Flouri, 2022, Fig. 9a-c), we note that the introgression event was not well supported by the DCT/BLT triplet tests of Suvorov *et al.* (2022, Fig. 3). Those tests supported introgression between *D. lowei* and *D. affinis*, and between *D. lowei* and *D. azteca*, but not between *D. lowei* and

D. athabasca (Table S3.1). Thus we replaced event iii by two events involving the daughter branches, between *D. lowei* and *D. affinis* (or $p \leftrightarrow q$), and between *D. lowei* and *D. azteca* (or $u \leftrightarrow v$) (Fig. 3.1a).

Our initial model of gene flow for clade 2 thus involved three gene-flow events: one ancestral and two involving extant taxa (Fig. 3.1a). As the triplet methods used by Suvorov *et al.* (2022) are agnostic about the direction of gene flow, we treated each event as a bidirectional gene-flow event. This way of determining the direction of gene flow involves a computational cost but was found to work well in simulations (Thawornwattana *et al.*, 2023a). We fitted both the MSC-I and MSC-M models of gene flow. Two variants of the MSC-I model were considered, which differed in the time order of the two introgression events involving *D. lowei* ($u \leftrightarrow v$ and $p \leftrightarrow q$). The results are summarized in Table 3.1 for MSC-I and Table 3.2 for MSC-M.

Under the MSC-I model, introgression from branches *ac* to *rb* (or $x \to y$, Fig. 3.1a) had the strongest signal. The estimated introgression probability was $\hat{\phi}_{x\to y} = 0.098$ and 0.092 for the two data halves, while the Bayes factor $B_{10} = \infty$ for the Bayesian test (Table 3.1, models 1a&1b). Introgression in the opposite direction ($y \to x$) was found to be absent, with the model of introgression rejected strongly ($B_{10} \leq 0.01$). Apart from the $x \to y$ introgression, all other introgression events were rejected at the $B_{10} \leq 0.01$ cut-off (Table 3.1). Note that the Bayesian test may strongly favor the null model and reject the more general model of gene flow, unlike hypothesis testing, which may fail to reject the null hypothesis but may never support it strongly.

The MSC-M model produced results consistent with the MSC-I model (Table 3.2, model 1). Similarly the only gene-flow event supported was from $x \to y$, with the estimated rate to be $M_{x\to y} =$ 0.32 and 0.34 migrants per generation for the two halves, respectively, while gene flow in the opposite direction was found to be absent. Also the $x \to y$ migration was the only one that was significant ($B_{10} > 100$), while all other gene-flow events were rejected by the Bayesian test at the $B_{10} \leq 0.01$ cut-off.

Interestingly the time of the $x \to y$ introgression under the MSC-I model was nearly identical to the divergence time at the mother node *a* (Fig. 3.1a): $\hat{\tau}_x = \hat{\tau}_y = 0.0261$ (with the 95% HPD CI 0.0257–0.0265) and 0.260 (0.0256–0.0265) for the two halves, respectively, compared with $\hat{\tau}_a = 0.0261$ (0.0257–0.0265) and 0.0261 (0.0256–0.0265) under models 1a and 1b of Table 3.1. This may suggest

that the introgression event was assigned to the wrong branch in the initial model; Huang *et al.* (2022a) found that when introgression is incorrectly assigned onto a daughter or mother branch of the lineage genuinely involved in gene flow, the introgression time tends to get stuck on the species divergence time. Thus we considered a model in which the $x \to y$ introgression was replaced by introgression involving the parental branch ($w \to z$). This model produced greater estimates of the introgression probability, $\varphi_{w\to z} = 0.248$ (CI 0.207–0.291) for the first half, and 0.393 (0.349–0.440) for the second half, and with the introgression time away from the species divergence time.

We also fitted an MSC-I model with both $w \to z$ and $x \to z$ introgressions (Fig. 3.1b), with the expectation that introgression event that did not occur should have low estimated rates, rejected by the test (Huang *et al.*, 2022a; Thawornwattana *et al.*, 2022). The analysis detected very strong evidence for gene flow between the sister lineages, with $\hat{\varphi}_{w\to z} = 0.728$ (0.689–0.769) and 0.712 (0.681–0.743) for the two data halves (Table 3.1, model 2). The evidence for the $x \to y$ introgression was also significant although the rate was much lower, at $\hat{\varphi}_{x\to y} = 0.069$ (0.055–0.083) and 0.069 (0.056–0.082) (Table 3.1, model 2).

We further assessed possible impacts of including an outgroup species, using either *D. melanogaster* or *D. insularis* as the outgroup, besides the 11 ingroup species in clade 2 (Tables S3.2 & S3.3). Some parameters such as the population size for the root of the species tree are known to be sensitive to the inclusion of outgroup species (Burgess and Yang, 2008). The introgression probabilities ($\varphi_{w\to z}, \varphi_{x\to y}$) and introgression times ($\tau_w = \tau_z, \tau_x = \tau_y$) are very similar among the datasets (for two halves and two outgroups) (Tables S3.2 & S3.3), and also similar to the estimates without the outgroup (Table S3.4). Estimates of θ_r varied depending on the outgroup used (Tables S3.2 & S3.3), possibly because branch r ancestral to clade 2 represents different populations depending on the outgroup.

Given the introgression events between extant species inferred using triplet summary methods (Suvorov *et al.*, 2022), we fitted MSC-I models incorporating various introgression events between extant species, when the $w \to z$ and $x \to y$ introgression events are already accommodated in the model (Table S3.5). In particular, we tested bidirectional introgression events involving *D. lowei* (Fig. 3.1a). All gene-flow events involving extant species, including bidirectional introgression events involving *D. lowei*, were rejected, with $B_{10} \leq 0.01$ (Table S3.5). The $w \to z$ and $x \to y$ introgressions remained

the only significant events, and parameter estimates were virtually identical to those under model 2 with the $w \to z$ and $x \to y$ introgressions only (Table S3.4). We examine the impact of taxon sampling on inference of gene flow below.

In the MSC-M model, we also included the $w \to z$ migration in addition to the $x \to y$ migration (Table 3.2, model 2). Similarly we obtained high estimates of migration rate between the sister lineages, $M_{w\to z} = 0.568$ (CI 0.518–0.615) and 0.603 (0.555–0.652) immigrants per generation, and the Bayesian test was highly significant. The migration rate for the non-sister lineages was much lower, estimated to be $\hat{M}_{x\to y} = 0.011$ and 0.009 for the two halves, and was not significant according to the test. Thus the evidence for the $x \to y$ gene flow was inconsistent between the MSC-I and MSC-M models. This could be due to weak signal or low information content in the data, or lower power of the MSC-M model than the MSC-I model (Thawornwattana *et al.*, 2024).

By integrating all analyses above, we suggest model 2 of Figure 3.1b as our final inferred model for clade 2 on the *Drosophila* phylogeny (Suvorov *et al.*, 2022), which includes both the $x \to y$ and $w \to z$ introgression events.

3.2.2 Estimation of model parameters on the *Drosophila* species tree

We fitted the final model of Figure 3.1b to estimate model parameters, with gene flow accommodated using either the MSC-I or the MSC-M models. Estimates of the rate of gene flow (φ in MSC-I and M in MSC-M) are given in Tables 3.1 & 3.2 (model 2), while those for all parameters are in Table S3.4.

As discussed in the section above, the estimated rate of gene flow between the sister lineages ($w \to z$) was very high under both the MSC-I and MSC-M models (Table 3.1, model 2; Table 3.2, model 2). In comparison, the estimated rate of $x \to y$ gene flow was much lower, and was indeed not significant under the MSC-M model. Here we ask whether the two models recover similar amounts of gene flow between the sister lineages. If the MSC-M model is the true model with the $w \to z$ migration occurring over a time period $\Delta\tau$, the expected cumulative proportion of migrants in the recipient population z will be

$$\varphi_0 = 1 - \mathrm{e}^{-4M_{wz}\Delta\tau/\theta_z} \tag{3.2}$$

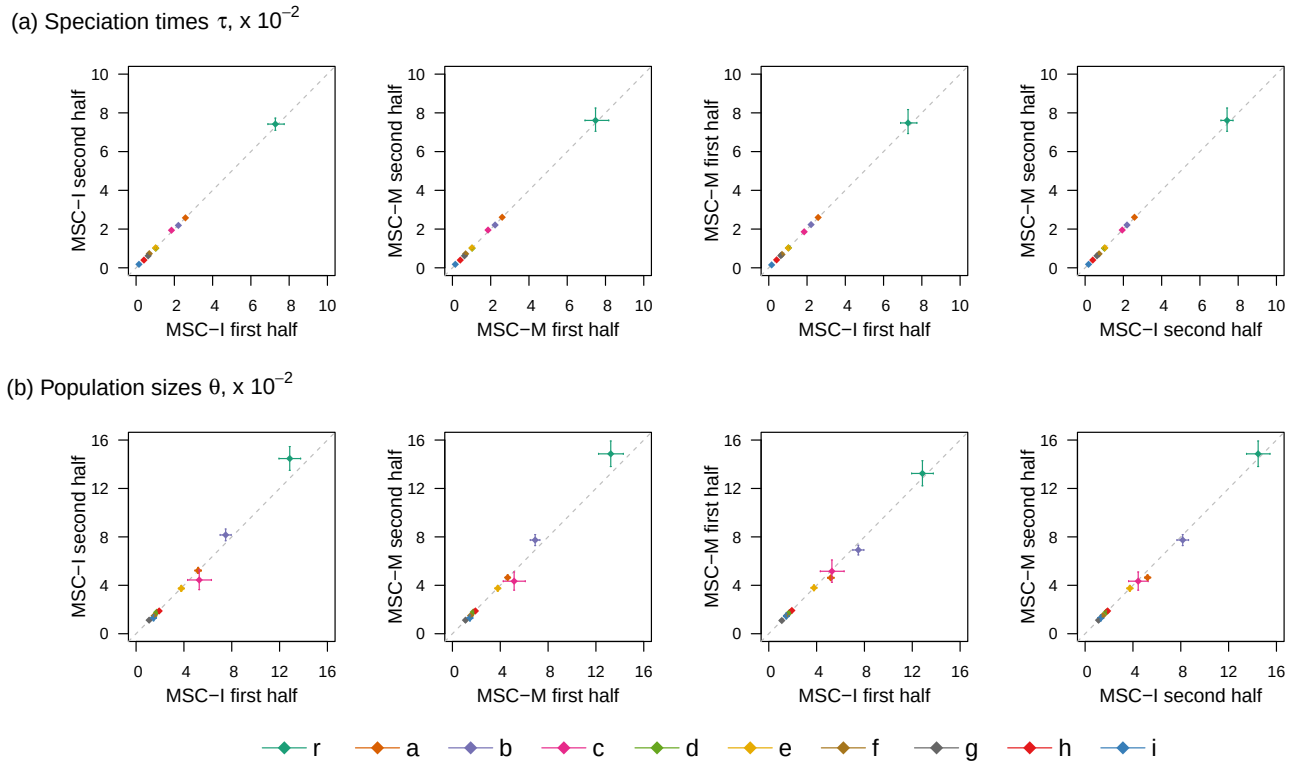

Figure 3.3: Posterior means and 95% HPD CIs for (**a**) species divergence times (τ, mutations per site) and (**b**) population sizes (θ) in the final model of Figure 3.1b obtained from BPP analyses of the *Drosophila* data under the MSC-I and MSC-M models.

(Huang *et al.*, 2022a). Using the estimates under MSC-M (Table S3.4), we calculated the expected introgression probability for the MSC-I model to be $\varphi_0 = 1 - e^{-4\times 0.568\times(0.0748-0.0260)/0.0692} = 0.798$ for the first half, and 0.789 for the second half, compared with the estimates under the MSC-I: 0.728 and 0.712. The estimates are similar, with slightly more gene flow inferred under MSC-M than under MSC-I.

Estimates of species divergence times (τ) and population sizes (θ) for the two data halves under the MSC-I and MSC-M models are shown in Figure 3.3. The four data-model combinations produced nearly identical estimates. Estimates of the age of the root for the clade (τ_r) differ considerably depending on whether gene flow is accommodated in the model (cf. Fig. 3.1a and Fig. 3.1b). This is consistent with previous studies which have shown that ignoring gene flow between species leads to serious underestimation of species split times (Leaché *et al.*, 2014; Thawornwattana *et al.*, 2023a; Tiley *et al.*, 2023).

For both data halves, $\hat{\varphi} > \frac{1}{2}$ under MSC-I, so that the majority of the lineages from descendent

species of node *a* (i.e., *D. subobscura, D. guanche, D. obscura, D. bifasciata*) are traced back to the introgression branch *rz* rather than the speciation branch *rw* (Fig. 3.1b). This is also the prediction of the MSC-M model since the estimates suggest $\varphi_0 > \frac{1}{2}$ by eq. 3.2. We also note that τ_r in model 1 (Fig. 3.1a) was similar to $\tau_w = \tau_z$ in model 2 (Fig. 3.1b). Thus the histories of sequence divergences reflected in the gene trees predicted by the two models (one with the $x \to y$ gene flow only and the other with both $x \to y$ and $w \to z$ gene flow) are somewhat similar.

Our analysis using BPP assumed the JC model (Jukes and Cantor, 1969). To see whether the mutation/substitution model affects the results, we analysed the data under the final MSC-I and MSC-M models of figure 3.1**b** assuming the GTR mutation model (Tavaré, 1986; Yang, 1994) instead of JC (Fig. S3.2, table S3.6). The estimates under the JC and GTR models were very similar, and the mutation model had little effects. Estimates of introgression probabilities and migration rates were also very similar between the two models (table S3.6). This robustness to the mutation model is expected because the main role of the mutation model in BPP analyses is to correct for multiple hits at the same site. As the sequence data from closely related species are extremely similar, any mutation model including the infinite-sites model (Takahata *et al.*, 1995) should work well. Similar observations were made by Shi and Yang (2018) and Flouri *et al.* (2022).

3.2.3 The impact of taxon sampling on inference of gene flow involving *D. lowei*

While there was significant evidence for gene flow between *D. lowei* and either *D. affinis* or *D. azteca* in the DCT/BLT tests of Suvorov *et al.* (2022, data S2), those gene-flow events were rejected in our analyses of data including all species in the group (Table 3.1). We thus examined the impact of taxon sampling, by constructing three triplet datasets and three quintet datasets and analyzing them using BPP. We focus on gene flow between *D. lowei* and *D. affinis*, for which the evidence was significant in two out of three triplets in the analysis of Suvorov *et al.* (2022, data S2).

First, we analysed the triplet datasets using QUIBL and DCT/BLT to examine the impact of the outgroup species. The assumed ingroup tree was ((*X*, *D. lowei*), *D. affinis*), where *X* was *D. pseudoobscura*, *D. persimilis*, or *D. miranda*, while *D. guanche* was used as the outgroup (Fig. 3.1a). Unrooted quartet trees were generated using RAXML under the JC model, rooted with the outgroup, and then

used as input for DCT/BLT and QuIBL. All summary-based tests were significant for all triplets. Suvorov *et al.* (2022) used *Anopheles gambiae* as the outgroup, and inferred quartet gene trees under the GTR+I+G model, finding that BLT and QuIBL were significant for all three triplets, while DCT was significant in two out of three triplets. The *Anopheles* outgroup is very distantly related to the ingroup species, and a closely related outgroup may be preferable as long as it is not involved in hybridization with the ingroup species. Nevertheless, the results from the summary methods are consistent between the two studies.

Next we analysed the triplet datasets using BPP (Table S3.7). Bidirectional introgression between *D. lowei* and *D. affinis* was specified in the MSC-I model. In all three datasets, there was strong evidence for introgression from *D. lowei* $\rightarrow$ *D. affinis*, with $B_{10} > 100$ and the estimated introgression probability $\hat{\varphi}_{p\rightarrow q} = 4.2$–$4.8\%$. There was also strong evidence rejecting introgression in the opposite direction (with $B_{10} \leq 0.01$ and $\hat{\varphi}_{q\rightarrow p} \approx 0.00$). Thus the BPP analysis of triplet datasets is consistent with the summary methods (DCT/BLT), although BPP was able to infer the direction and strength of gene flow, rejecting the $q \rightarrow p$ introgression.

Finally the quintet datasets which include two outgroup species, *D. guanche* and *D. obscura*, were analysed using BPP under MSC-I assuming bidirectional introgression between *D. lowei* and *D. affinis*, either with or without accommodating the $w \rightarrow z$ and $x \rightarrow y$ introgressions (Fig. 3.1b, Table S3.7). In all cases, the $q \rightarrow p$ introgression was rejected, as in the analysis of the triplet data. Without the $w \rightarrow z$ and $x \rightarrow y$ introgressions in the model, the $p \rightarrow q$ introgression rate was low (1-2%) and was not significant (with $B_{10} < 100$ in all three datasets). When the $w \rightarrow z$ and $x \rightarrow y$ introgressions were assumed in the model, the $p \rightarrow q$ introgression became significant in all three quintet datasets (with $B_{10} > 100$), with $\hat{\varphi}_{p\rightarrow q} \approx 4.1$–$5.7\%$ (Table S3.7). We also note that in the analysis of data from all 11 species in clade 2, under the model which incorporates the $w \rightarrow z$ and $x \rightarrow y$ introgressions, the estimated introgression probability $\varphi_{p\rightarrow q}$ was very low (0.1–0.2%) and was rejected with $B_{10} \leq 0.01$ (Table S3.5, last section).

In summary, while the $q \rightarrow p$ introgression was rejected in all analyses, the Bayesian test of the $p \rightarrow q$ introgression was sensitive to the species included in the data and to whether other major introgression events ($w \rightarrow z$, $x \rightarrow y$) were already accounted for in the model. The reasons for this

sensitivity are not well-understood. We suspect that part of the difficulty may be due to problems of sampling, as the data consist of only one sequence per species per locus. The introgression probability is defined as a proportion of migrants in the recipient species. Knowledge of the population size or genetic diversity of the recipient species should help our inference of the contribution to that diversity from introgression. We note that the population size parameters $\theta_{D.\ lower} = \theta_p$ and $\theta_{D.\ affinis} = \theta_q$ are very poorly estimated with wide CIs, and the introgression probability $\varphi_{p\to q}$, if nonzero, was relatively low ($< 6\%$) (Table S3.7), so that inference may be easily affected by factors other than gene flow. Including multiple samples per species may be expected to increase the information in the data about the $p \to q$ introgression (see Discussion).

3.2.4 Analyses of simulated data by Bayesian and summary methods: Drosophila-based simulation

Our Bayesian analysis of the *Drosophila* clade-2 data of Suvorov *et al.* (2022) produced different results from those obtained by Suvorov *et al.* (2022) using triplet methods. To understand possible reasons for the differences, we conducted two sets of simulations to study the statistical behaviors of the methods.

In the *Drosophila*-based simulation, we used parameter estimates of Tables S3.2 (first half) & S3.3 (first half) obtained from the BPP analysis of the clade-2 data including the *D. insularis* outgroup under the final MSC-I and MSC-M models with the $w \to z$ and $x \to y$ gene-flow events. Two data halves, each of 1388 loci, were simulated. Bayesian estimates of parameters (Table S3.8) were very close to the true parameter values, and the 95% HPD CIs were similar to those in the analysis of the real data (cf: Table S3.4).

In the BPP analyses, we used diffuse gamma priors on parameters τ and θ with the prior means matching the true values (the 1x priors): $\tau_r \sim \text{G}(2,50)$ and $\theta \sim \text{G}(2,200)$. To assess the impact of the priors, we varied the prior means to be either 10 times larger (the 10x priors): $\tau_r \sim \text{G}(2,5)$ and $\theta \sim \text{G}(2,20)$, or 10 times smaller (the 0.1x priors): $\tau_r \sim \text{G}(2,500)$ and $\theta \sim \text{G}(2,2000)$. The priors had little impact on estimation of the species split times, but some population size parameters were somewhat affected, with the use of the 0.1x priors causing underestimation of θ_r and θ_c (Fig. S3.3).

Estimates of introgression probabilities (φ) and migration rates (M) were very close to the true values (Table S3.9). Overall the posterior was robust to such orders-of-magnitude changes to the prior mean, apparently because the datasets analysed in this study were large.

Note that the major introgression event in the true model, from $w \to z$, is between sister lineages and is thus unidentifiable by triplet methods used by Suvorov *et al.* (2022). Instead we applied DCT (which is based on gene-tree counts) and BLT (which is based on branch lengths) to detect the $x \to y$ introgression by constructing triplets. In 8/28 triplets significant evidence was detected by DCT. No signal was detected by BLT.

3.2.5 Analyses of simulated data by Bayesian and summary methods: quartet data

In the second set of simulations, we used the MSC-M and MSC-I models for four species (A, B, C, and outgroup O) of Figure 3.2, with gene flow between non-sister lineages (B, C). Divergence times (τ) and population sizes (θ) resemble estimates from the *Drosophila* data, but we used a range of parameter values. Each dataset was analysed using BPP under both the MSC-M and MSC-I models, resulting in eight simulation-analysis settings. We examine both estimation of model parameters (in particular the rate of gene flow) and Bayesian test for the presence of gene flow. This set of simulation is similar to previous studies that examined the properties of the Bayesian method (Huang *et al.*, 2020, 2022a; Thawornwattana *et al.*, 2023a, 2024), but here we included a number of summary methods.

Bayesian estimation in quartet data. Here we discuss estimation of the rate of gene flow (φ in MSC-I and M in MSC-M) (Fig. 3.4).

In the M-M and I-I settings (Fig. 3.4), data were simulated and analysed under the same model. The rate of gene flow was well estimated, with the posterior means around the true values while the 95% HPD CIs become narrower when the data size increases. In informative datasets, the coverage of the 95% CI was in general $> 95\%$. Introgression probability was more precisely estimated in the inflow model (with gene flow from $C \to B$, Fig. 3.2a&b) than in the outflow model ($B \to C$, Fig. 3.2c&d) (Fig. S3.6 inflow I-I vs. Fig. S3.10 outflow I-I). For example, the CI width in the least informative data set ($L = 250, S = 2, n = 250$, $\theta = 0.0025$) was $\sim$43% narrower under the inflow than outflow models. These results are consistent with the observation of Thawornwattana *et al.* (2023a).

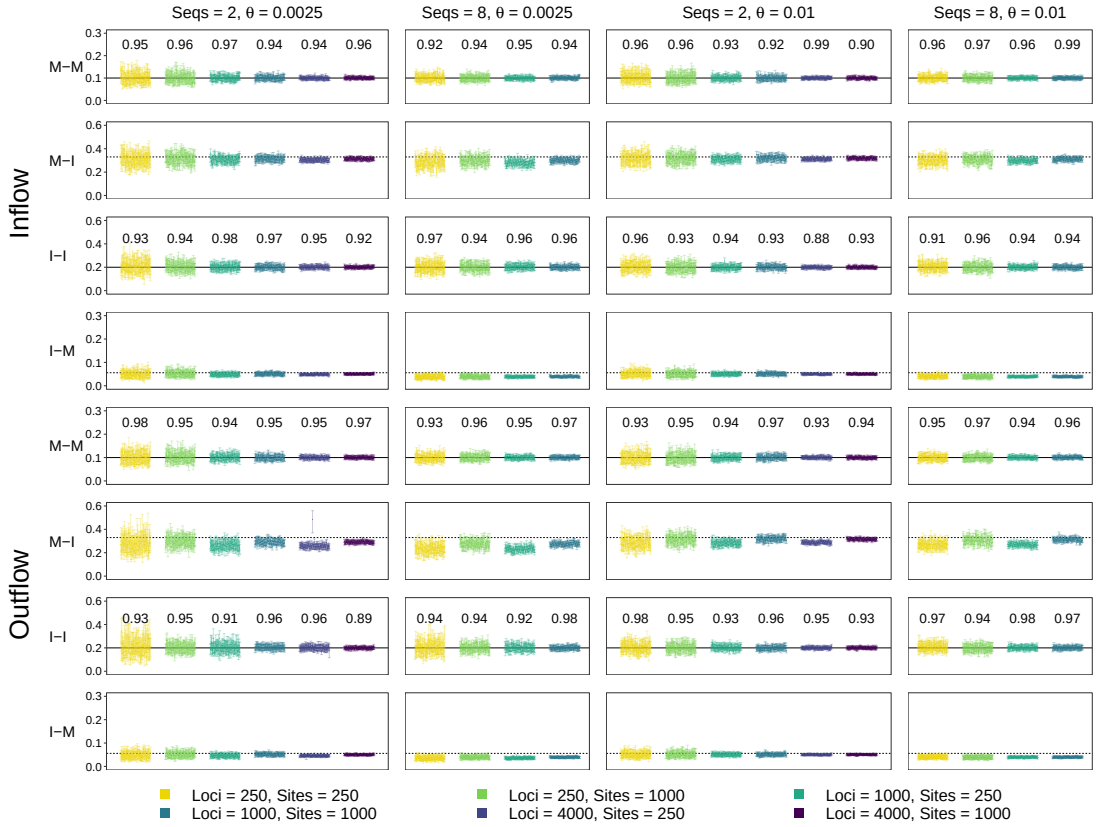

Figure 3.4: Posterior means and 95% CIs for introgression probabilities (φ) in the MSC-I model and migration rates (M) in the MSC-M model obtained from the BPP analysis of 100 simulated data replicates. Datasets were simulated under the four models of Figure 3.2 and analysed under both the MSC-M and MSC-I models, with eight settings in total. For example, in the inflow-M-I setting, replicate datasets were simulated under the inflow-migration (MSC-M) model (Fig. 3.2**a**) and analysed under the introgression (MSC-I) model (Fig. 3.2**b**). Results for other parameters in the eight simulation settings are in Figures S3.4–S3.11. Numbers above the CI bars represent the CI coverage probability. Solid black lines represent true parameter values. Dashed black lines represent the theoretical expectations when the mode of gene flow is misspecified (eq. 3.2). Large datasets under settings with $L = 4000$ loci and $S = 8$ sequences per species per locus (with either 250 or 1000 sites) were not analysed.

In the M-I and I-M settings (Fig. 3.4), the mode of gene flow was misspecified. The analysis of Huang *et al.* (2022a) suggests that when data are generated under MSC-M but analysed under MSC-I, not all gene flow that has occurred is recoverable, with $\hat{\varphi} < \varphi_0$ (eq. 3.2). This was the case in the simulation here (Figs. 3.4, inflow M-I and outflow M-I). The underestimation was more serious (with larger difference between $\hat{\varphi}$ and φ_0) in the outflow case than in the inflow case.

The results for all parameters summarized in Figures S3.4–S3.11. Estimates of population sizes (θ) and divergence/introgression times (τ) are shown there.

In the M-M settings (Figs. S3.4 & S3.8), data were simulated under MSC-M and analysed under

the same model so that the correct model of gene flow was specified. Similarly under the I-I settings (Figs. S3.6 & S3.10), the MSC-I model was used to both simulate and analyse data. These represent the best-case scenarios and provide a reference for comparison. In all cases, species divergence times (τ_R, τ_S, τ_T) were very well estimated, as were population sizes for extant species ($\theta_A, \theta_B, \theta_C, \theta_O$). In the I-I setting, the introgression time ($\tau_X = \tau_Y$ in Fig. 3.2b & d) was well-estimated as well. The posterior means approach the true values while the 95% HPD CIs become narrower when the data size increases. The results were consistent with the asymptotic expectation that the CI width should reduce by a half when the number of loci quadruples ($L = 250$ versus 1000) (O'Hagan and Forster, 2004, p.73). In informative datasets, the coverage of the 95% CI was in general higher than 95%. Ancestral population sizes ($\theta_R, \theta_S, \theta_T$) had larger uncertainties, especially at the low mutation rate ($\theta = 0.0025$) and in small datasets (with 250 loci and 250 sites), as did φ in MSC-I or M in MSC-M. Note that in our simulation, species divergence times (τs) are proportional to θ so that the two values of θ mimic the use of genome regions with different mutation rates. The patterns are consistent with previous simulations conducted under the MSC and MSC-I models (Huang *et al.*, 2020). Introgression probability was more precisely estimated in the inflow model (with gene flow from the outgroup species C to the ingroup species B, Fig. 3.2a&b) than in the outflow model (with gene flow from the ingroup species B to the outgroup species C, Fig. 3.2c&d) (Fig. S3.6 inflow I-I vs. Fig. S3.10 outflow I-I). For example, the CI width in the least informative data set ($L = 250, S = 2, n = 250$, $\theta = 0.0025$) was ~43% narrower under the inflow than outflow models. These results are consistent with the observation of Thawornwattana *et al.* (2023a).

In the M-I settings (Figs. S3.5 for inflow-M-I & S3.9 for outflow-M-I), data were simulated under MSC-M and analysed under MSC-I, so that the mode of gene flow was misspecified. The analysis of Huang *et al.* (2022a) suggests that when data are generated under MSC-M but analysed under MSC-I, not all gene flow that has occurred is recoverable, and the estimated amount under MSC-I ($\hat{\varphi}$) is less than the expected amount (φ_0) given by eq. 3.2, with $\hat{\varphi} < \varphi_0$. This was indeed the case in the simulation here (Figs. 3.4, inflow M-I and outflow M-I). The underestimation was more serious (with larger difference between $\hat{\varphi}$ and φ_0) in the outflow case than in the inflow case, and for short sequences ($n = 250$) than for long sequences ($n = 1000$). While gene flow occurs throughout the time period

$(0, \tau_T)$ (Fig. 3.2a&c), the estimated introgression time ($\hat{\tau}_X$) was smaller than the mid-time ($\tau_T/2$) and closer to the end of the time period (0), in particular, for datasets with many sequences per species ($S = 8$ versus 2) and for long sequences ($n = 1000$ versus 250) (Figs. S3.5 & S3.9). This is because the estimated introgression time is dominated by the most recent sequence divergence time between species (Huang *et al.*, 2022a). Among parameters present in both the MSC-M and MSC-I models, species divergence times were accurately estimated except for τ_S which showed a small negative bias (Figs. S3.5 & S3.9). Population sizes for extant and ancestral species were well-estimated as well, although there seemed to be a small positive bias in θ_S. Overall estimates of shared parameters were very similar to those in the I-I setting where there was no model misspecification.

In the I-M settings (Figs. S3.7 & S3.11), data were simulated under MSC-I and analysed under MSC-M. The estimated migration rate under MSC-M ($\hat{M}$) was less than the predicted rate under the true MSC-I model (M_0), with $\hat{M} < M_0$ (Figs. S3.7 for inflow-I-M and Fig. S3.11 for outflow-I-M). Thus not all gene flow that occurred according to the true MSC-I model was recovered by the misspecified MSC-M model. Again, parameters common in both models, including species divergence times and population sizes for extant and ancestral species, were accurately estimated with little bias (Figs. S3.7 & S3.11). There was no discernible difference from estimates in the M-M setting in which there was no model misspecification.

Whilst the MSC-M and MSC-I models make very different assumptions about the mode of gene flow, in the simulation settings examined here (Figs. S3.4–S3.11), both produced reliable estimates of divergence times and population sizes even when the mode of gene flow was misspecified.

Bayesian test in quartet data. Bayesian test of gene flow overall showed very high power in simulated quartet data (Fig. 3.5). At the 1% cut-off (i.e, with $B_{10} > 100$), the test achieved ~100% power in all simulation settings. This was the case even in the least informative datasets (with $L = 250$ loci, $n = 250$ sites, and at the low mutation rate with $\theta = 0.0025$). In particular, power was ~100% in the M-I and I-M settings as well, when the mode of gene flow was misspecified. For instance, if gene flow occurred continuously over an extended time period according to the MSC-M model but was assumed to occur in a pulse in the MSC-I model, the test still detected gene flow with nearly full power (Fig. 3.5, inflow-M & outflow-M, BPP-wrong model).

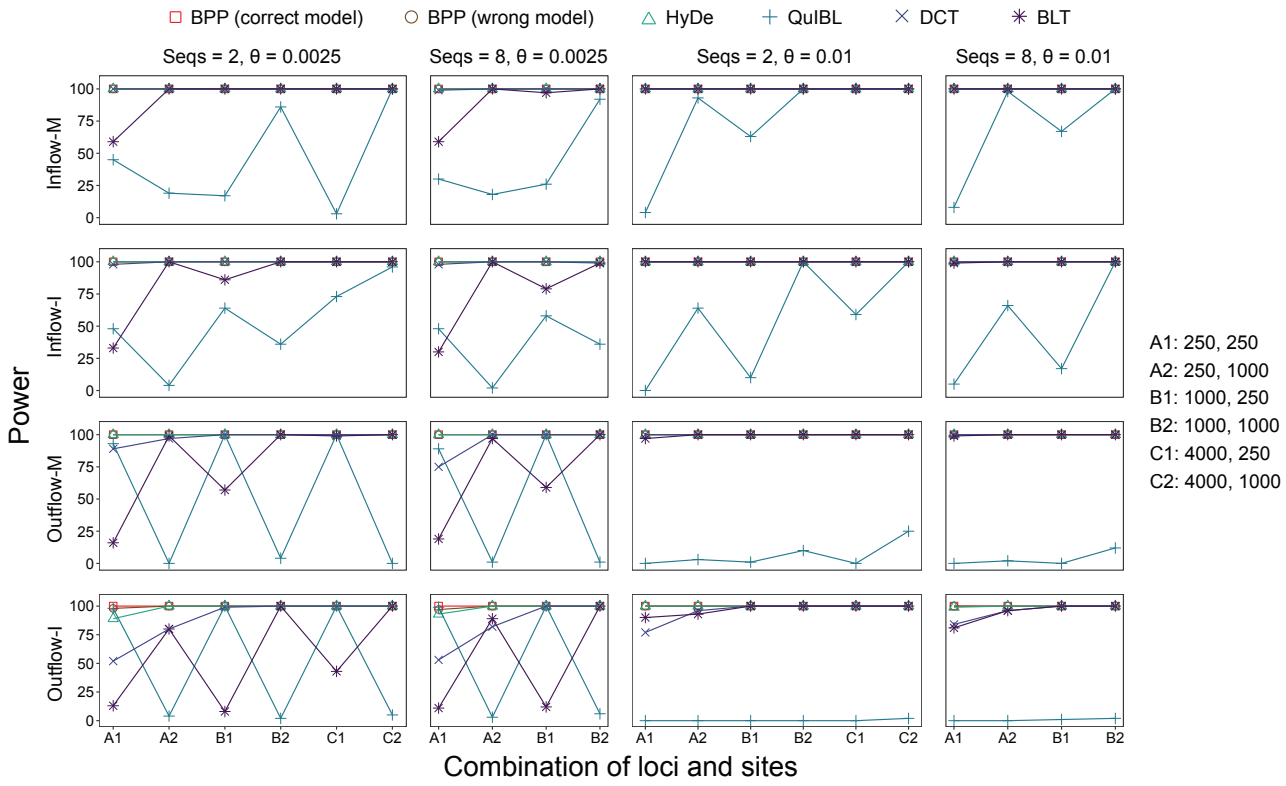

Figure 3.5: Power (percentage of replicates in which the null model of no gene flow is rejected at the 1% level) of BPP (MSC-I and MSC-M), HyDe, QuIBL, DCT and BLT to detect gene flow in data simulated under the four models of gene flow in Figure 3.2. Bayesian test of gene flow using BPP is conducted assuming either the correct model (e.g., Inflow-M-M) or incorrect model (e.g., Inflow-M-I), with gene flow detected if the Bayes factor $B_{10} > 100$. Data configurations are specified in the number of loci (L) and the number of sites (n): for example, in configuration "A2: 250, 1000", each dataset consists of $L = 250$ loci, each of $n = 1000$ sites. Bayesian estimates of parameters from the same data are shown in Figures 3.4 and S3.4–S3.11.

Estimation by summary methods in quartet data. We applied several summary methods to estimate the introgression probability (Fig. 3.6) and to test for gene flow (Fig. 3.5). For data simulated under MSC-I (Fig. 3.6, inflow-I and outflow-I), all summary methods for estimating φ appeared to be biased. In the inflow scenario, SNaQ and HyDe overestimated the introgression probability, while DCT and QuIBL produced underestimates (Fig. 3.6, inflow-I). In the outflow scenario, all summary methods produced underestimates (Fig. 3.6, outflow-I). QuIBL, in particular, produced gross underestimates. This bias of the QuIBL method was noted previously by Edelman *et al.* (2019).

Test of gene flow by summary methods in quartet data. Next we examined the power of summary methods for testing for gene flow, in comparison with BPP (Fig. 3.5). While BPP achieved ~100% power in all datasets, even when the mode of gene flow was misspecified, the performance of the summary methods varied. The two methods based on gene-tree branch lengths, QuIBL and BLT, had

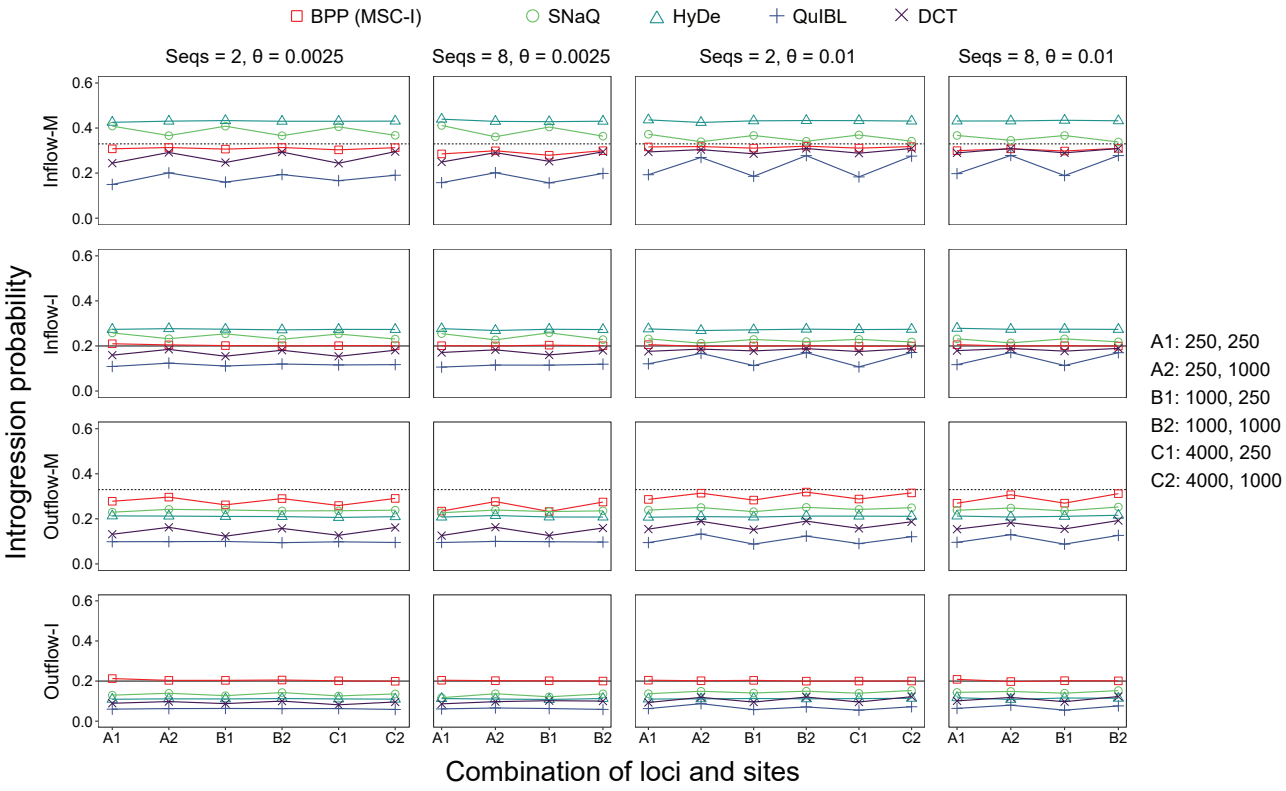

Figure 3.6: Average estimates of introgression probability (φ) produced by BPP (MSC-I), SNaQ, HyDe, QuIBL and DCT for each of the four gene-flow models of Figure 3.2. Black solid lines represent the true value of φ in the MSC-I model, whereas dashed lines represent the expected value φ_0 (eq. 3.2) when the data are generated under the MSC-M model. See legend to Figure 3.5.

particularly low power for short sequences (250 sites instead of 1000) and at the low mutation rate (with $\theta = 0.0025$ instead of 0.01). This may be expected since short and highly similar sequences contain little phylogenetic information, leading to large sampling errors in the estimated branch lengths, while those errors are ignored by both methods. QuIBL had ~0% power in data generated under the outflow model. This appeared to be due to the fact that QuIBL assumes a triplet species tree with an inflow model of introgression rather than outflow (Edelman *et al.*, 2019, Figs. S61&S62), so that for those data the assumed direction of gene flow was incorrect.

HyDe showed good power. As it uses site-pattern counts pooled over loci, it is not sensitive to sampling errors in the estimated gene-tree topology and branch lengths at each locus. We note that HyDe is based on a hybrid-speciation model, which is a special case of the inflow model with symmetry in the population size (Ji *et al.*, 2023). Previously HyDe was found to perform poorly when those assumptions were not met; in particular, HyDe was found to lack power when gene

flow occurred in the opposite (outflow) direction (Ji *et al.*, 2023, Figure 9; Pang and Zhang, 2024). In the simulation here the method performed relatively well (Fig. 3.5), apparently because the same population size was usd for all species in the simulation model, so that the assumptions of HYDE were largely met.

Finally DCT showed low power in the least informative datasets, but was not so sensitive to short sequences as were QUIBL and BLT (Fig. 3.5). This may be because DCT uses gene-tree topologies but not branch lengths.

3.3 Discussion

3.3.1 Likelihood and summary methods for inferring gene flow

Our simulations highlight the desirable statistical properties of the Bayesian method implemented in BPP. The power to detect gene flow via the Bayesian test (Ji *et al.*, 2023) was high, even when the information content of the dataset was low and even if the mode of gene flow was misspecified. Bayesian estimation of parameters including introgression probabilities and migration rates was highly accurate. We found that if the mode of gene flow was misspecified (when the true model was MSC-I and the analysis model was MSC-M, or vice versa), the Bayesian method may underestimate the amount of gene flow. However, the shared parameters between the two models were reliably estimated. The simulation results here are consistent with and extend previous simulations which examined the Frequentist properties of Bayesian test and Bayesian estimation under the MSC model with gene flow (Huang *et al.*, 2020, 2022a; Ji *et al.*, 2023; Pang and Zhang, 2024; Thawornwattana *et al.*, 2023a).

The performance of summary methods in the simulation varied considerably (Figs. 3.5 & 3.6). All summary methods for estimating the introgression probability were found to be biased (Fig. 3.6). In particular, branch length-based methods such as QUIBL performed poorly, and had low power to detect gene flow, except in the most informative inflow datasets. When the species are closely related and the sequences are highly similar, estimated branch lengths in reconstructed gene trees are expected to have considerable errors and uncertainties, which may affect the performance of those

methods.

Suvorov *et al.* (2022) has relied on the f-branch approach to integrate results of many triplet analyses. This was designed to move introgression events to ancestral branches on the species tree, as gene flow involving ancestral lineages may show up as significant introgression events in many species triplets, which may be hard to interpret (Malinsky *et al.*, 2018). Disturbingly a recent study demonstrated that the commonly used triplet methods, such as the *D*-statistic, HYDE, and SNAQ, do not have the ability to identify different introgression models, including ancestral introgression from an outgroup, and inflow and outflow between non-sister lineages (Pang and Zhang, 2024). It is unclear how the performance of f_{branch} is affected by such unidentifiability. In general, research is needed to understand the behavior of the approach in realistic scenarios involving multiple introgression events on a species tree of more than three species when test of gene flow is always conducted using species triplets.

Overall analyses of real and simulated data in this study as well as in previous studies (Huang *et al.*, 2020, 2022a; Ji *et al.*, 2023; Pang and Zhang, 2024; Thawornwattana *et al.*, 2023a) have highlighted large gaps between full likelihood methods (such as BPP) and summary methods. Summary methods are orders-of-magnitude faster computationally and can easily accommodate genome-scale datasets, while likelihood methods have much better statistical performance (with higher power in inferring gene flow and less bias in estimating its rate). There is an urgent need for improving the statistical properties of summary methods and the computational efficiency of likelihood methods for inferring gene flow using genomic sequence data.

3.3.2 Gene flow in *Drosophila*

There has been long-standing interest in gene flow between species on the *Drosophila* phylogeny. Noor *et al.* (2000) analysed within-species polymorphism and between-species divergence along the genome to infer gene flow between *D. pseudoobscura* and *D. persimilis*. The population genetic analysis did not identify the direction of gene flow. Wang and Hey (2010, see also Dalquen *et al.*, 2017) explicitly modelled the coalescent-with-migration process in the so-called isolation-with-migration (IM) model and used multilocus sequence data to infer low but significant gene flow from *D. sim-*

ulans to *D. melanogaster*, with no gene flow in the opposite direction. The study of Suvorov *et al.* (2022) is noteworthy for its use of 155 *Drosophila* genome assemblies, covering the whole *Drosophila* genus and suggesting multiple instances of between-species gene flow.

Our re-analysis of data for clade 2 in the *Drosophila* genus of Suvorov *et al.* (2022) has confirmed the authors' overall conclusion that gene flow is prevalent on the species phylogeny, and extended that work by characterizing the lineages involved in gene flow and its direction and by estimating the timing and rates of gene flow. We inferred a gene-flow event involving sister lineages which is unidentifiable by the triplet summary methods used by Suvorov *et al.* (2022) while some introgression events inferred by Suvorov *et al.* (2022) were rejected in our Bayesian test. Our simulation in general demonstrates the accuracy and robustness of BPP, and raised concerns about the reliability of the summary methods used by Suvorov *et al.* (2022). Our analyses suggest a need for a re-analysis of gene flow for the other clades on the *Drosophila* phylogeny.

Here we note a few limitations with both our Bayesian analysis and the sequence data, which may affect our inference. First, our search in the space of models was not exhaustive. We used the Bayesian test to confirm or remove gene-flow events proposed in the triplet analyses of Suvorov *et al.* (2022), and in some cases repositioned events to ancestral branches when our analysis suggested incorrect placement (Table 3.1). We also assessed various scenarios of gene flow involving extant species (Table S3.5). This constitutes a limited search in the space of introgression models. The use of a stringent cut-off for B_{10} in the test may lead to false negatives (i.e., failure to detect gene flow when it exists), but the test appeared to be very powerful in simulations (this study and Ji *et al.*, 2023).

Second, some concerns may be raised about the suitability of the sequence data of Suvorov *et al.* (2022). The data consist of single-copy protein-coding genes compiled to infer the phylogeny and to estimate divergence times for the whole *Drosophila* genus, with divergence times >50MY (or >100MY from the *Anopheles gambiae* outgroup). While single-copy orthologous genes are ideal for phylogenetic reconstruction and divergence time estimation among distantly related species, which are major objectives of the study of Suvorov *et al.* (2022), they may not be optimal for inferring gene flow between closely related species. The data for clade 2 involve a high degree of incompleteness, with missing species at ∼50% of the gene loci. Noncoding parts of the genome tend to have

higher mutation rates and may be more informative than conserved exons, even though they may pose challenges to genome assembly. Also the data appear to be "haploid consensus sequences", with genotypic phase at heterozygous sites in the diploid sequence resolved effectively at random, creating chimeric sequences that may not exist in nature and may impact on genealogy-based analyses under the MSC (Andermann *et al.*, 2019; Huang *et al.*, 2022b). Furthermore, the data consist of only one sample per species per locus. Summary methods considered here do not use information in multiple samples per species, and indeed some authors suggest that "adding more samples provides little new information with respect to introgression" (Hibbins and Hahn, 2022). However, likelihood-based methods such as BPP can accommodate multiple samples per species, and both theoretical analysis and computer simulation suggest that including multiple samples per species (in particular for species receiving immigrants) may boost the information content in the data for inferring gene flow (Huang *et al.*, 2020; Yang and Flouri, 2022). For example, with one sequence per species, some models of introgression are unidentifiable but the problem disappears when multiple samples are included in the data (Thawornwattana *et al.*, 2023a; Yang and Flouri, 2022). It is unclear whether the extreme sensitivity in the inference of the *D. lowei* $\rightarrow$ *D. affinis* ($p \rightarrow q$) introgression to taxon sampling (Table S3.7) is due to the joint effects of the use of one sample per species and the 'pseudohaploidization' of the haploid consensus sequences, as the 'unusualness' of the chimeric sequences from the ingroup species may depend on inclusion or exclusion of sequences from more distant species. Note that haploid consensus sequences may be chimeric sequences that do not exist in natural populations and may thus appear highly unusual. They may show up on gene trees as long branches or deeply divergent lineages, and may thus affect inference methods such as BPP that are based on gene genealogies (Huang *et al.*, 2022b, Fig. 6, Table 6).

While issues related to data quality may impact our analyses using BPP, the major introgression event involving sister lineages inferred in our analysis (Fig. 3.1b) appears to be robust and well supported. However, it is likely that certain instances of gene flow may be missed in our analyses. We leave it to future studies to assemble sequence datasets including noncoding parts of the genome and including multiple samples per species to infer gene flow in this group of species. In this regard we note that (Kim *et al.*, 2023) has discussed the complexities of *Drosophila* genome assembly and made

progress in producing high-quality genomic data.

3.4 Supplemental Information

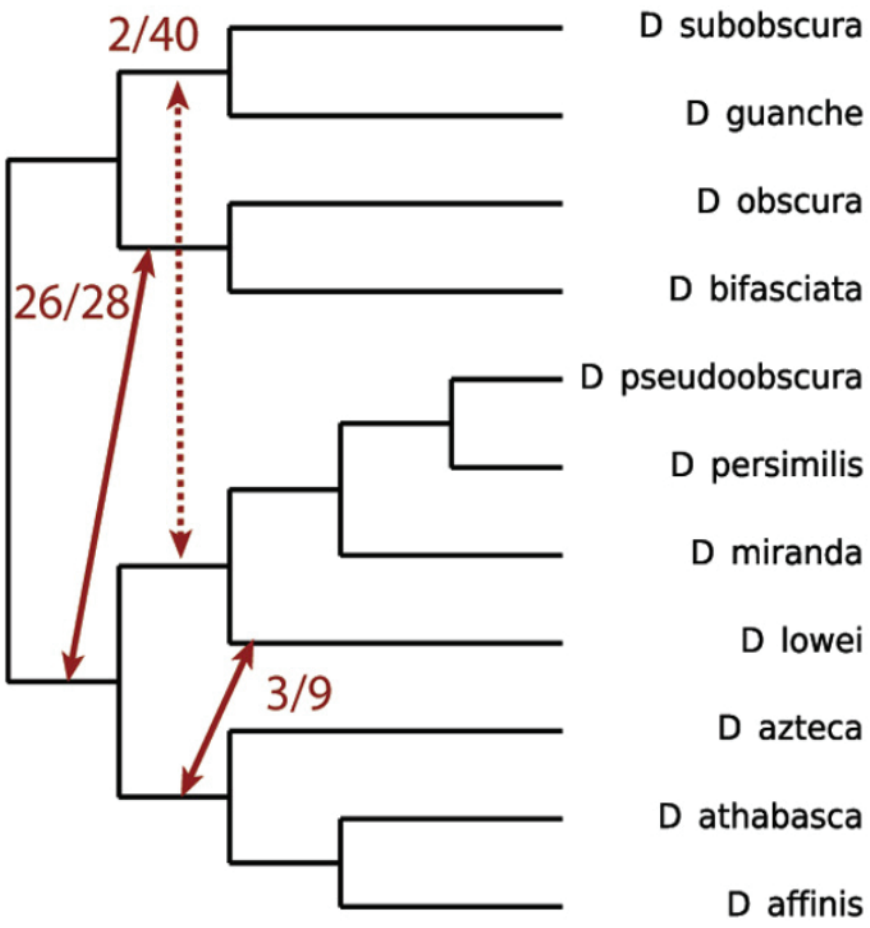

Figure S3.1: Model of gene flow for clade 2 inferred by Suvorov et al. (2022, Fig. 3). The fractions next to each arrow represent the number of species triplets that support the introgression event by both DCT and BLT out of the number of informative triplets tested (table S3.1).

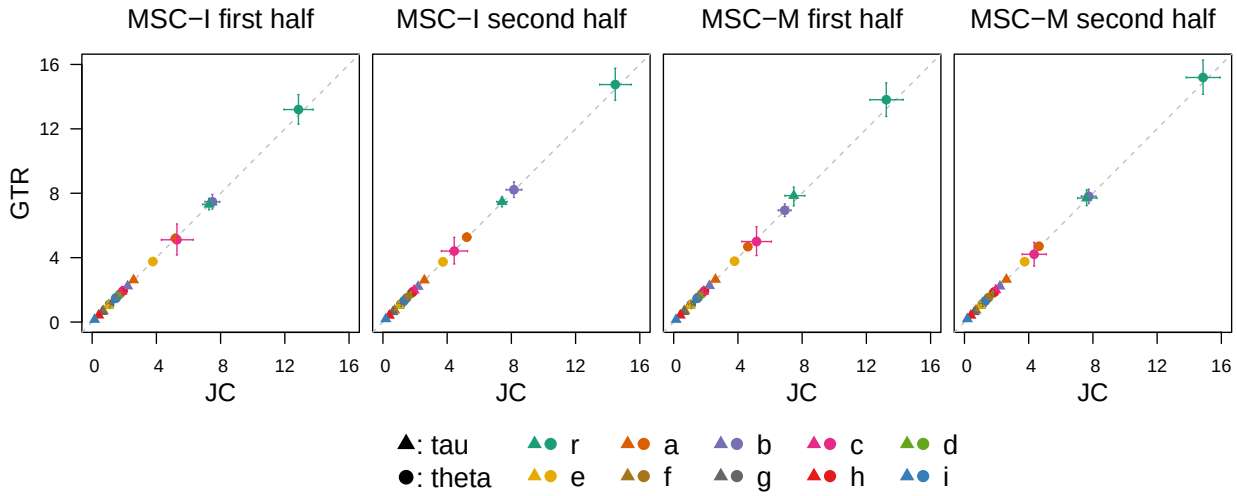

Figure S3.2: Posterior means (points) and 95% HPD CIs (bars) for τ and θs in the final model of figure 3.1**b** in BPP analysis of the *Drosophila* data under the JC (Jukes and Cantor, 1969) and GTR (Tavaré, 1986; Yang, 1994) mutation models.

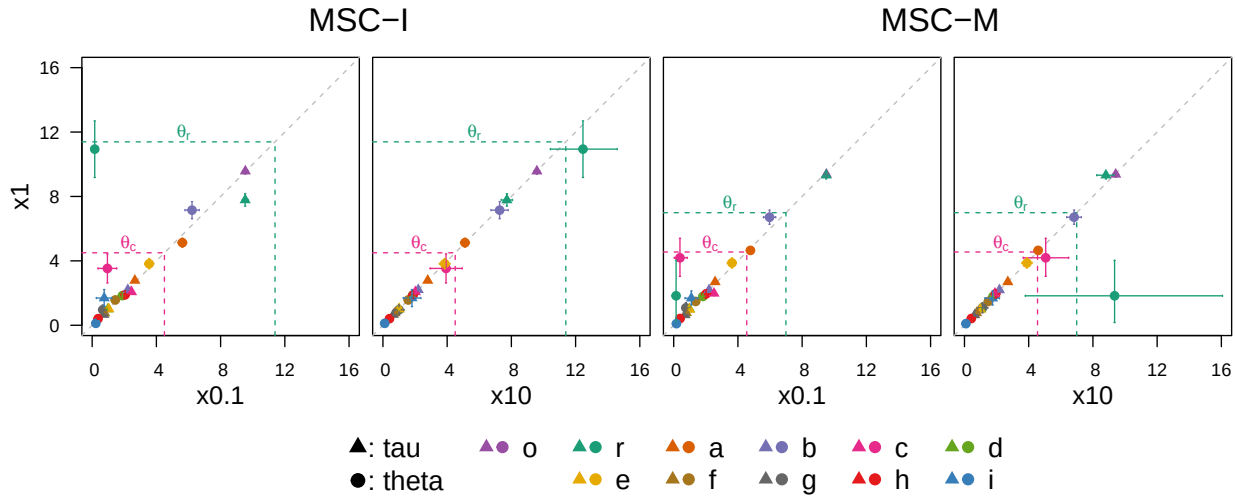

Figure S3.3: The impact of priors on τ and θ in BPP analysis of the data simulated using parameter estimates from the *Drosophila* data (Tables S3.2 & S3.3, *D. insularis* outgroup, first half). We used three sets of priors on τ and θ, with the prior means to be (i) equal to the true values (the 1x prior): $\tau_r \sim \text{G}(2,50)$ and $\theta \sim \text{G}(2,200)$; (ii) 10x smaller than the true values (the 0.1x prior): $\tau_r \sim \text{G}(2,500)$ and $\theta \sim \text{G}(2,2000)$; or (iii) 10x larger (the 10x prior): $\tau_r \sim \text{G}(2,5)$ and $\theta \sim \text{G}(2,20)$. Estimates of introgression probabilities and migration rates are in table S3.9. Estimates for all parameters under the 1x prior are in table S3.8.

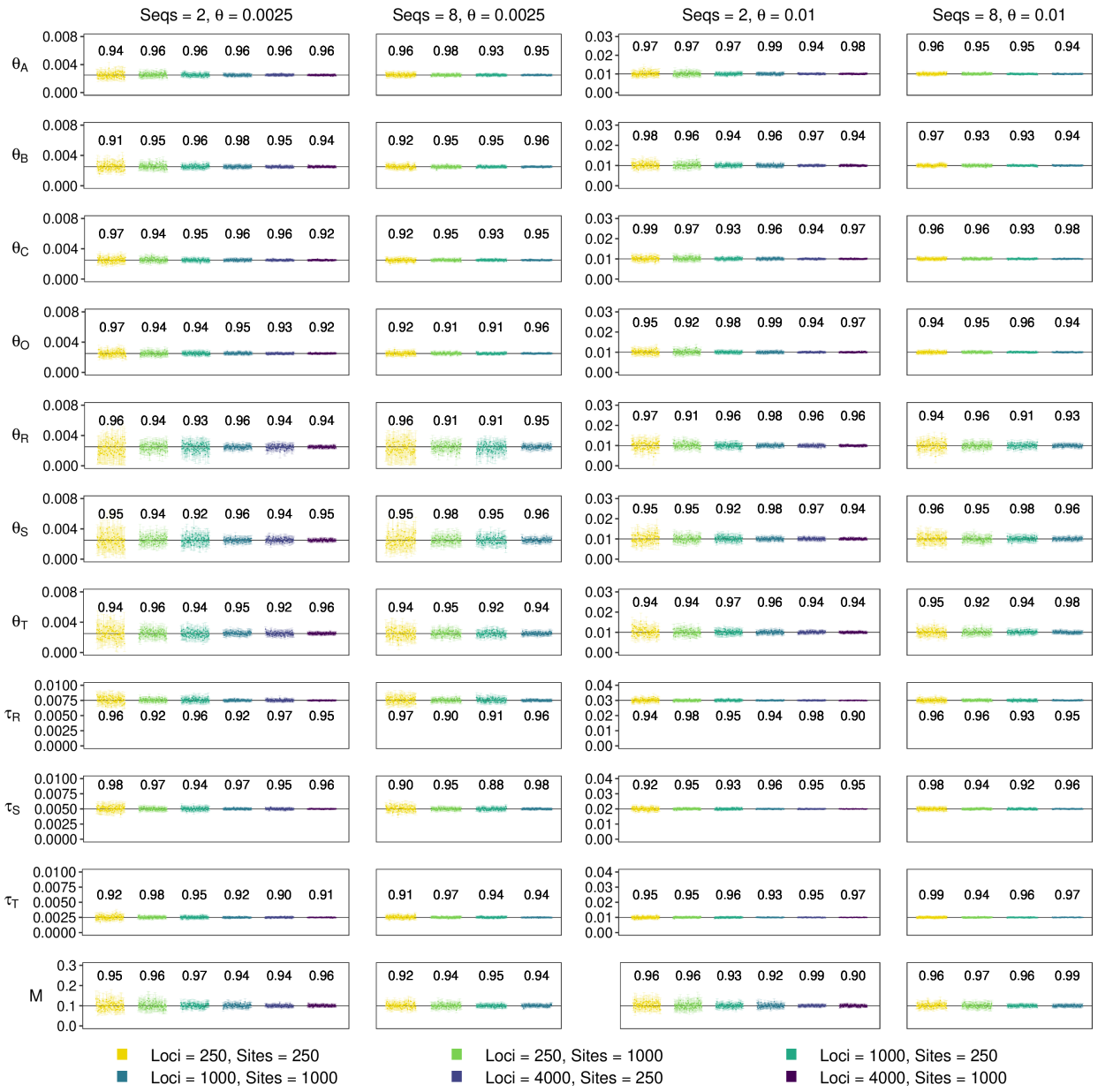

Figure S3.4: **(inflow-M-M)** Posterior means and 95% HPD CIs for parameters in BPP analysis of replicate datasets simulated and analysed under the inflow migration (MSC-M) model of figure 3.2a. Solid black lines represent true values. Numbers above (or below) the CI bars are the coverage probability. We use the 'simulation-analysis' format to specify our simulation setting, so that 'M-M' means that data were both simulated and analysed under the MSC-M model, while 'M-I' (Fig. S3.5) means that data were simulated under the MSC-M model and analysed under the MSC-I model.

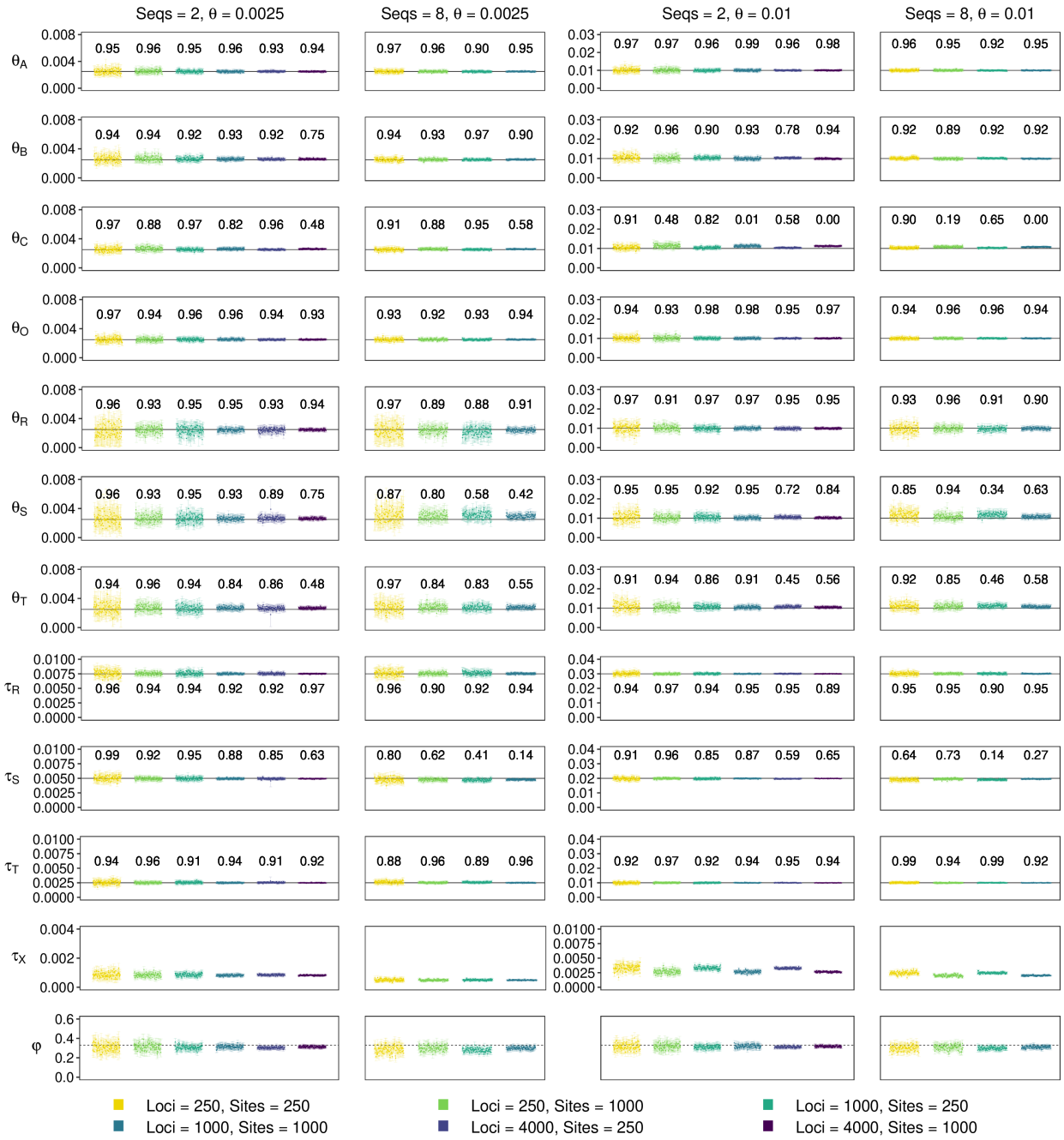

Figure S3.5: **(inflow-M-I)** Posterior means and 95% HPD CIs for parameters in BPP analysis of datasets simulated under the inflow migration ('M' for MSC-M) model (Fig. 3.2**a**) and analysed under the introgression ('I' for MSC-I) model (Fig. 3.2**b**). Dashed black lines for the φ parameter denote the theoretical value given by eq. 3.2. See legend to figure S3.4.

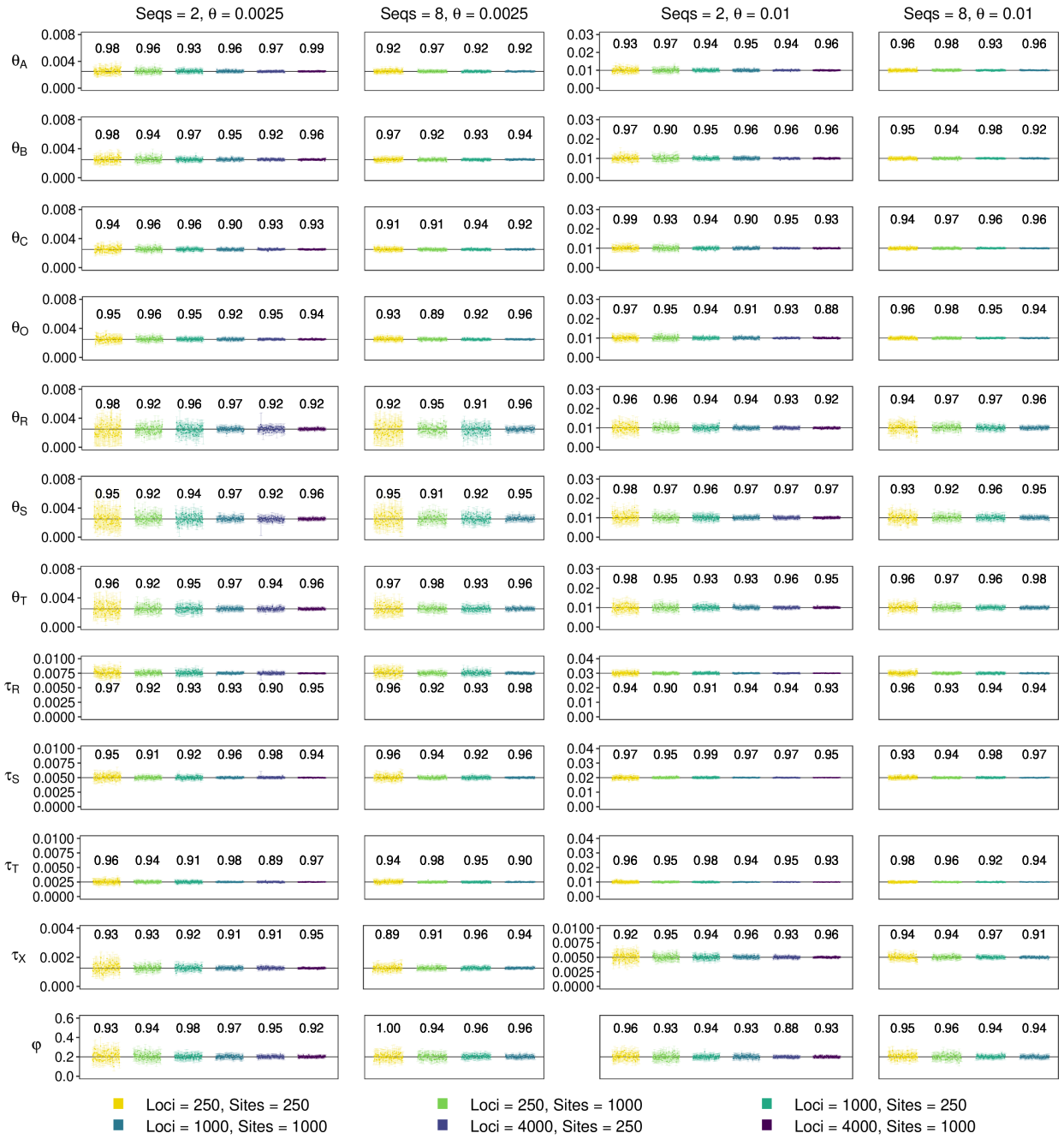

Figure S3.6: **(inflow-I-I)**. See legends to figures S3.4&S3.5.

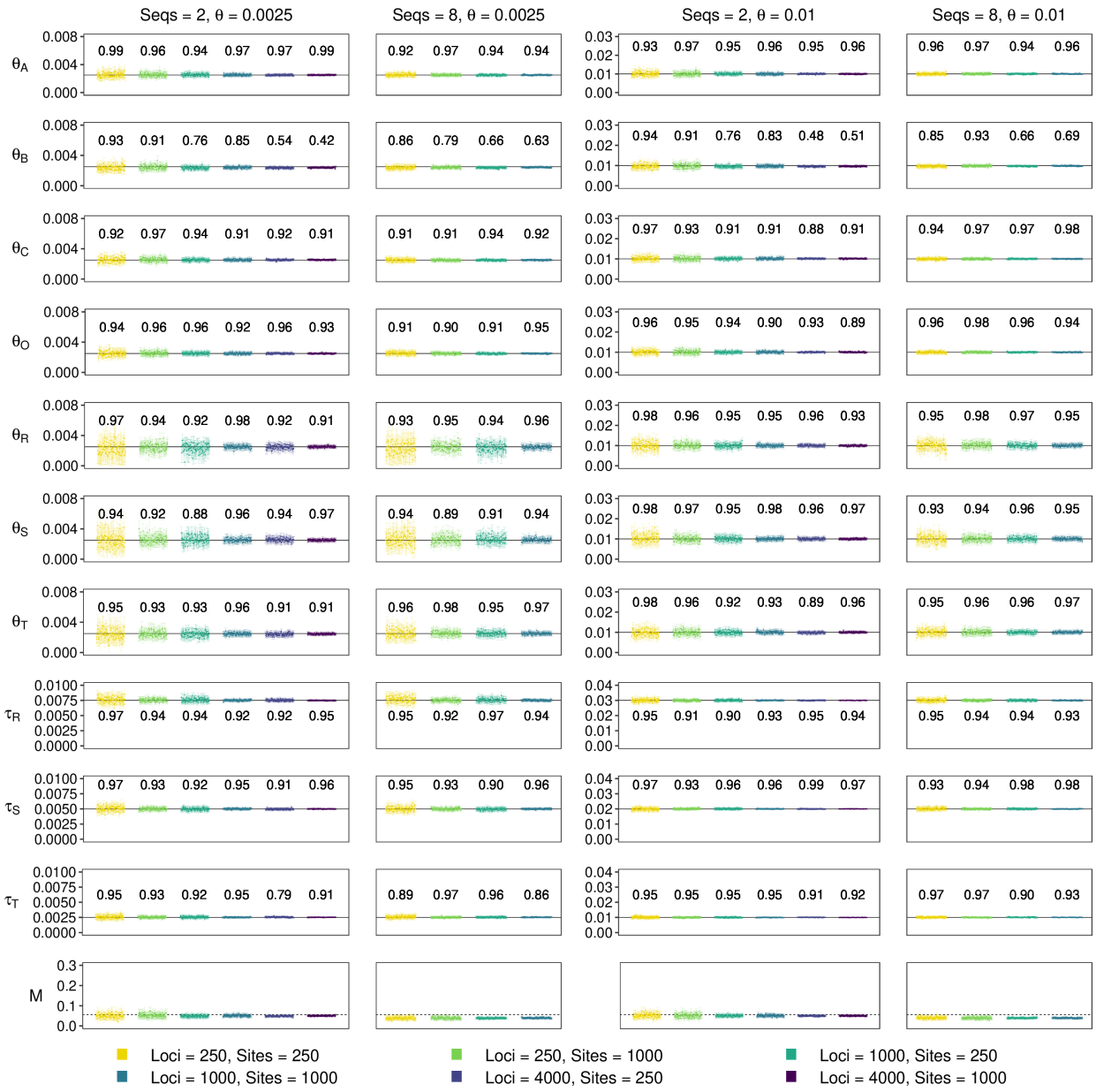

Figure S3.7: **(inflow-I-M)**. See legends to figures S3.4&S3.5.

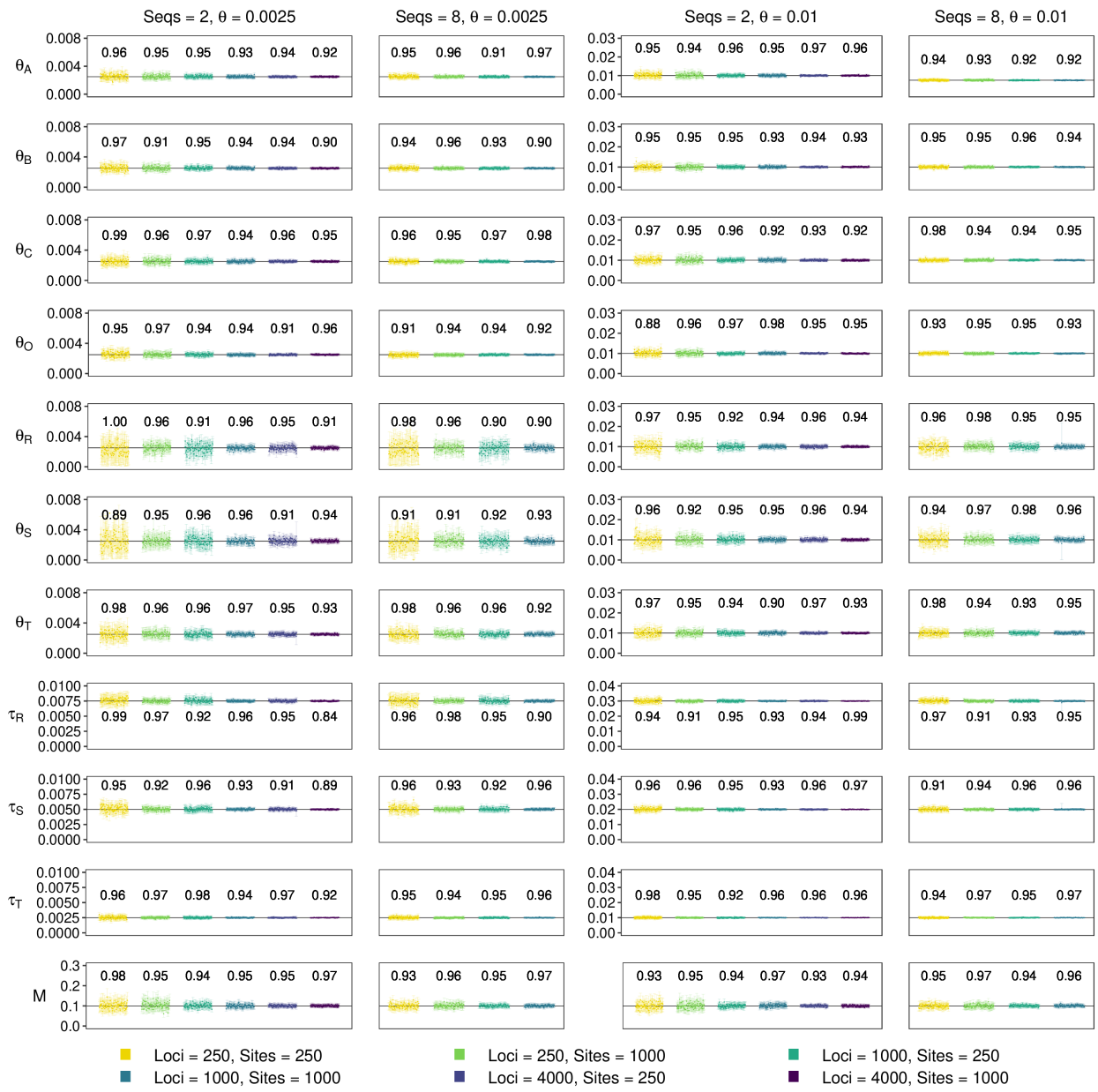

Figure S3.8: **(outflow-M-M)**. See legends to figures S3.4&S3.5.

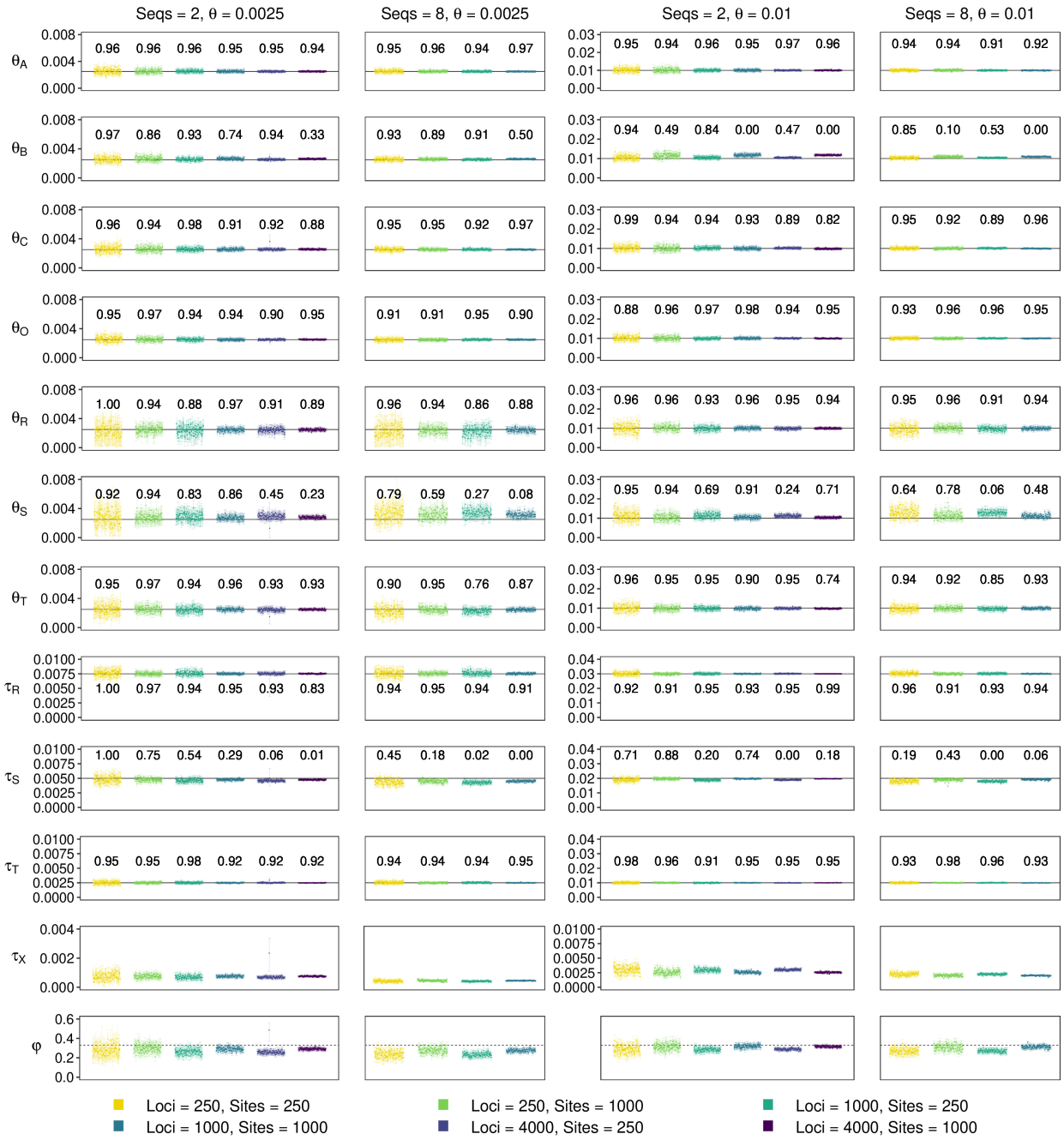

Figure S3.9: **(outflow-M-I)**. See legends to figures S3.4&S3.5.

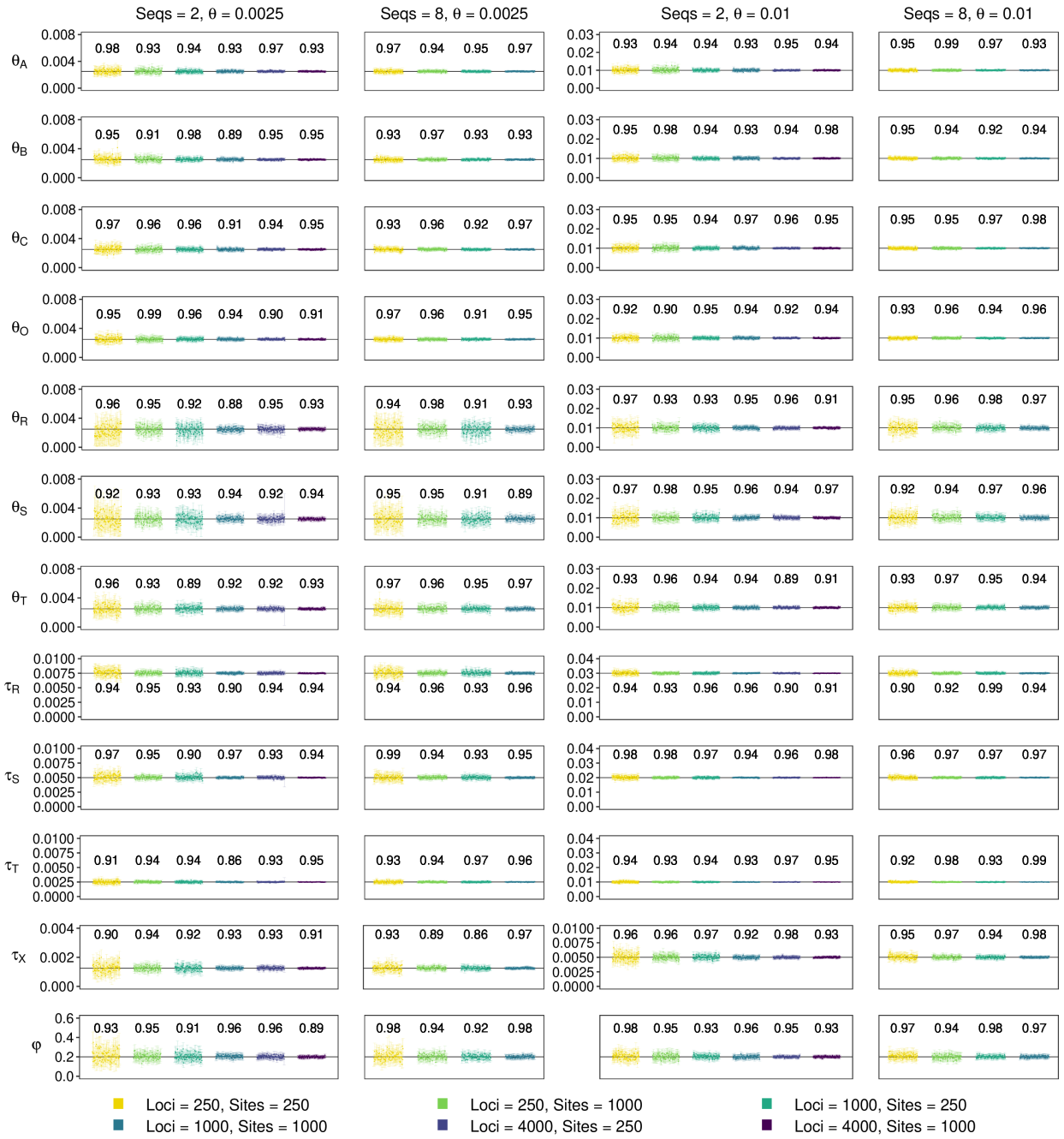

Figure S3.10: **(outflow-I-I)**. See legends to figures S3.4&S3.5.

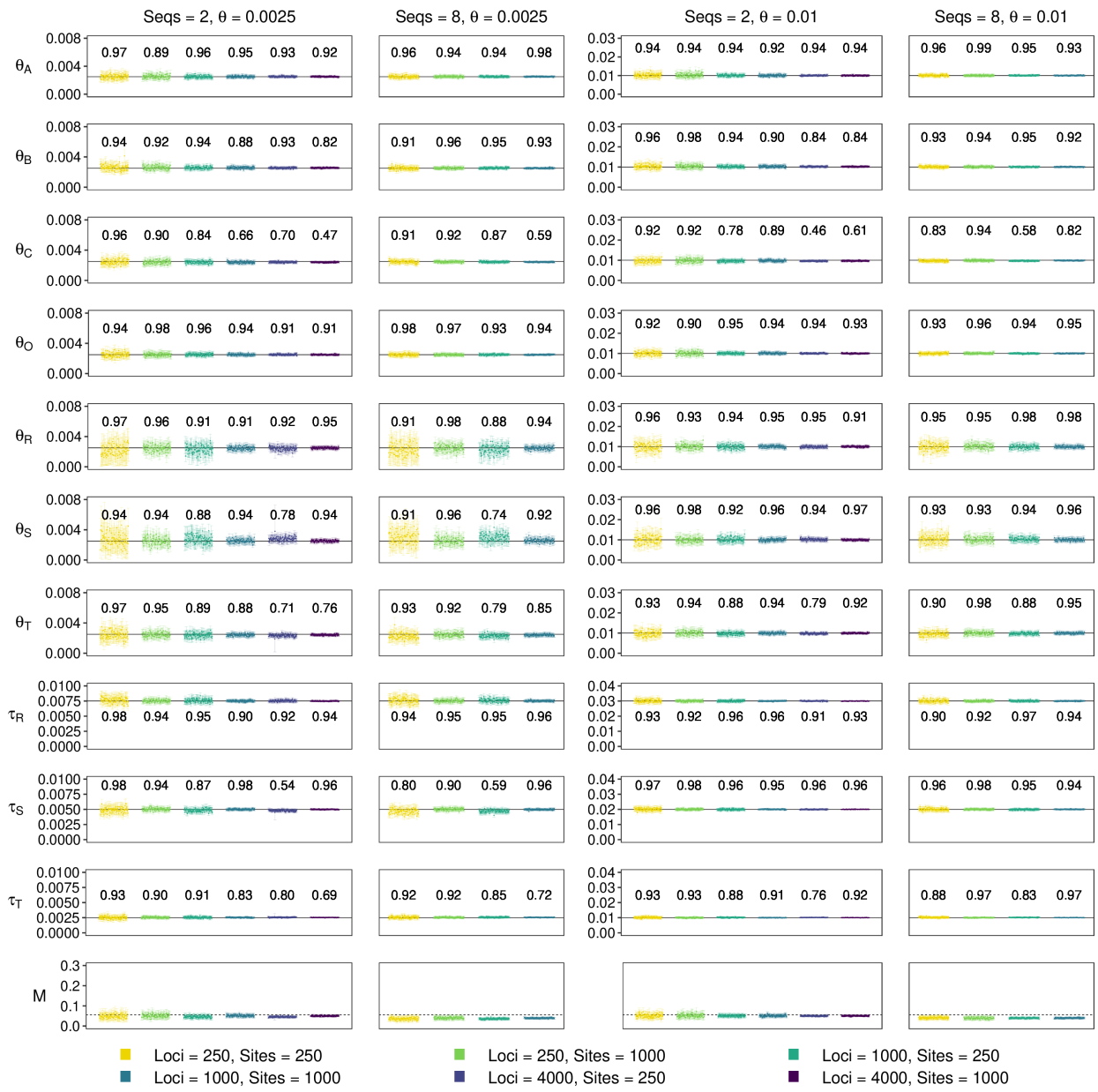

Figure S3.11: **(outflow-I-M)**. See legends to figures S3.4&S3.5.

Table S3.1: Fraction of species triplets that supported gene flow out of all informative triplets tested using DCT/BLT for the 11 *Drosophila* species of clade 2 (from Suvorov et al. 2022, Supplementary Data S2)

	subobscura	*guanche*	*obscura*	*bifasciata*	*pseudoobscura*	*persimilis*	*miranda*	*lowei*	*azteca*	*athabasca*
D. subobscura										
D. guanche	–									
D. obscura	0/1	0/1								
D. bifasciata	–	–	–							
D. pseudoobscura	0/3	0/3	2/5	2/2						
D. persimilis	0/3	0/3	2/4	–	–					
D. miranda	0/4	0/4	2/8	2/6	0/1	–				
D. lowei	1/6	1/5	3/9	3/8	–	–	0/2			
D. azteca	0/1	0/1	2/3	2/2	0/3	0/1	0/2	1/4		
D. athabasca	0/3	0/2	2/4	2/4	0/2	0/1	0/4	0/4	0/1	
D. affinis	–	–	2/3	2/2	0/1	–	0/2	2/3	–	–

Table S3.2: The impact of outgroup. Posterior means and 95% HPD CIs for parameters under the MSC-I models obtained from BPP analyses of the *Drosophila* data including an outgroup

	Outgroup = *D. insularis*		Outgroup = *D. melanogaster*	
Parameters	First half	Second half	First half	Second half
Population sizes (θ, $\times 10^{-2}$)				
θ_o	17.95 (17.08, 18.84)	17.99 (17.09, 18.92)	16.96 (16.21, 17.72)	17.15 (16.37, 17.94)
θ_r	11.39 (10.05, 12.73)	11.87 (10.70, 13.07)	1.38 (0.16, 2.91)	1.77 (0.12, 4.02)
θ_a	5.14 (4.89, 5.39)	5.22 (4.97, 5.47)	5.06 (4.83, 5.29)	5.26 (5.02, 5.50)
θ_b	7.20 (6.79, 7.61)	7.86 (7.41, 8.31)	7.01 (6.61, 7.40)	7.80 (7.37, 8.24)
θ_c	4.50 (3.73, 5.27)	3.65 (3.06, 4.26)	4.35 (3.65, 5.06)	3.95 (3.32, 4.61)
θ_d	1.72 (1.62, 1.83)	1.78 (1.67, 1.88)	1.74 (1.64, 1.84)	1.78 (1.67, 1.89)
θ_e	3.72 (3.49, 3.96)	3.69 (3.45, 3.93)	3.70 (3.47, 3.94)	3.71 (3.47, 3.95)
θ_f	1.57 (1.46, 1.67)	1.57 (1.47, 1.68)	1.58 (1.48, 1.68)	1.58 (1.47, 1.68)
θ_g	1.11 (0.97, 1.25)	1.11 (0.97, 1.26)	1.11 (0.97, 1.25)	1.13 (0.98, 1.28)
θ_h	1.92 (1.78, 2.06)	1.89 (1.75, 2.03)	1.92 (1.78, 2.07)	1.90 (1.75, 2.03)
θ_i	1.46 (1.22, 1.69)	1.31 (1.08, 1.56)	1.48 (1.25, 1.72)	1.33 (1.09, 1.58)
θ_w	$= \theta_a$	$= \theta_a$	$= \theta_a$	$= \theta_a$
θ_z	$= \theta_b$	$= \theta_b$	$= \theta_b$	$= \theta_b$
θ_x	$= \theta_c$	$= \theta_c$	$= \theta_c$	$= \theta_c$
θ_y	$= \theta_b$	$= \theta_b$	$= \theta_b$	$= \theta_b$
Speciation/introgression times (τ, $\times 10^{-2}$)				
τ_o	9.54 (9.33, 9.76)	10.05 (9.84, 10.26)	7.48 (7.35, 7.61)	7.90 (7.76, 8.04)
τ_r	7.53 (7.14, 8.00)	7.43 (7.09, 7.82)	7.42 (7.27, 7.57)	7.84 (7.67, 8.02)
τ_a	2.75 (2.71, 2.78)	2.74 (2.70, 2.78)	2.78 (2.74, 2.82)	2.78 (2.74, 2.82)
τ_b	2.18 (2.15, 2.21)	2.17 (2.14, 2.20)	2.23 (2.19, 2.26)	2.18 (2.15, 2.22)
τ_c	1.96 (1.89, 2.04)	2.07 (2.01, 2.13)	1.98 (1.91, 2.05)	2.06 (1.99, 2.12)
τ_d	1.03 (1.00, 1.05)	1.01 (0.99, 1.03)	1.03 (1.01, 1.05)	1.01 (0.99, 1.03)
τ_e	1.01 (0.98, 1.04)	1.03 (1.00, 1.06)	1.02 (0.99, 1.05)	1.03 (1.00, 1.06)
τ_f	0.75 (0.72, 0.77)	0.77 (0.74, 0.79)	0.75 (0.72, 0.77)	0.77 (0.75, 0.80)
τ_g	0.63 (0.60, 0.65)	0.63 (0.60, 0.65)	0.63 (0.61, 0.66)	0.62 (0.60, 0.65)
τ_h	0.41 (0.39, 0.43)	0.40 (0.38, 0.42)	0.42 (0.40, 0.43)	0.40 (0.38, 0.42)
τ_i	0.15 (0.13, 0.17)	0.18 (0.16, 0.21)	0.15 (0.13, 0.17)	0.18 (0.16, 0.20)
$\tau_w = \tau_z = \tau_{w \to z}$	4.00 (3.92, 4.09)	3.98 (3.92, 4.03)	3.98 (3.89, 4.07)	4.05 (3.99, 4.11)
$\tau_x = \tau_y = \tau_{x \to y}$	2.73 (2.69, 2.77)	2.72 (2.68, 2.77)	2.77 (2.72, 2.81)	2.77 (2.72, 2.81)
Introgression probabilities				
$ra \to rb$ (or $w \to z$)	0.708 (0.670, 0.747)	0.677 (0.640, 0.718)	0.672 (0.641, 0.703)	0.697 (0.665, 0.730)
$ac \to rb$ (or $x \to y$)	0.098 (0.080, 0.117)	0.080 (0.066, 0.094)	0.085 (0.067, 0.105)	0.083 (0.068, 0.097)

Note.– The outgroup species is either *D. insularis* or *D. melanogaster*, with the latter being a closer outgroup for clade 2 (Suvorov *et al.*, 2022, Fig. 1). Node r is the root for clade 2 (Fig. 3.1b), and o is the root of the tree including the outgroup. Note that *D. insularis* is a more distant outgroup to clade 2 (with a larger τ_o) than is *D. melanogaster*. The estimates for the first half with the *D. insularis* outgroup are used to simulate data analysed in table S3.8.

Table S3.3: The impact of the outgroup. Posterior means and 95% HPD CIs (in parentheses) under the MSC-M models obtained from BPP analyses of *Drosophila* data including an outgroup

	Outgroup = *D. insularis*		Outgroup = *D. melanogaster*	
Parameters	First half	Second half	First half	Second half
Population sizes (θ, $\times 10^{-2}$)				
θ_o	18.06 (17.18, 19.01)	17.97 (17.07, 18.88)	16.66 (15.93, 17.43)	16.77 (15.98, 17.52)
θ_r	6.99 (0.26, 13.79)	12.71 (10.94, 14.55)	1.44 (0.15, 3.09)	1.51 (0.18, 3.25)
θ_a	4.81 (4.58, 5.04)	4.81 (4.60, 5.02)	4.65 (4.44, 4.85)	4.74 (4.54, 4.94)
θ_b	6.83 (6.44, 7.20)	7.59 (7.18, 7.99)	6.57 (6.22, 6.93)	7.36 (6.98, 7.76)
θ_c	4.55 (3.81, 5.35)	3.59 (3.04, 4.20)	4.27 (3.64, 4.98)	3.87 (3.28, 4.51)
θ_d	1.73 (1.63, 1.84)	1.78 (1.68, 1.89)	1.75 (1.65, 1.86)	1.79 (1.68, 1.90)
θ_e	3.72 (3.48, 3.95)	3.68 (3.44, 3.92)	3.71 (3.49, 3.95)	3.72 (3.49, 3.96)
θ_f	1.58 (1.48, 1.68)	1.58 (1.47, 1.68)	1.59 (1.48, 1.69)	1.58 (1.47, 1.68)
θ_g	1.10 (0.97, 1.25)	1.11 (0.97, 1.26)	1.10 (0.97, 1.24)	1.12 (0.98, 1.27)
θ_h	1.92 (1.78, 2.06)	1.88 (1.75, 2.02)	1.92 (1.78, 2.06)	1.89 (1.75, 2.03)
θ_i	1.46 (1.23, 1.69)	1.31 (1.08, 1.56)	1.47 (1.25, 1.72)	1.33 (1.09, 1.58)
Speciation/introgression times (τ, $\times 10^{-2}$)				
τ_o	9.51 (9.29, 9.73)	10.05 (9.85, 10.26)	7.61 (7.49, 7.75)	8.11 (7.97, 8.25)
τ_r	8.77 (7.75, 9.69)	8.34 (7.90, 8.74)	7.60 (7.46, 7.74)	8.10 (7.95, 8.24)
τ_a	2.74 (2.70, 2.78)	2.75 (2.71, 2.79)	2.80 (2.77, 2.84)	2.80 (2.76, 2.84)
τ_b	2.20 (2.17, 2.23)	2.18 (2.15, 2.22)	2.24 (2.21, 2.28)	2.20 (2.16, 2.24)
τ_c	1.95 (1.87, 2.02)	2.08 (2.02, 2.14)	1.99 (1.92, 2.05)	2.07 (2.00, 2.13)
τ_d	1.02 (1.00, 1.04)	1.01 (0.99, 1.03)	1.02 (1.00, 1.05)	1.01 (0.99, 1.03)
τ_e	1.01 (0.98, 1.04)	1.03 (1.00, 1.05)	1.02 (0.99, 1.05)	1.03 (1.00, 1.06)
τ_f	0.74 (0.72, 0.77)	0.77 (0.74, 0.79)	0.74 (0.72, 0.77)	0.77 (0.75, 0.80)
τ_g	0.63 (0.61, 0.66)	0.62 (0.60, 0.65)	0.63 (0.61, 0.66)	0.62 (0.60, 0.65)
τ_h	0.41 (0.39, 0.43)	0.40 (0.38, 0.42)	0.42 (0.40, 0.43)	0.40 (0.38, 0.42)
τ_i	0.15 (0.13, 0.17)	0.18 (0.16, 0.20)	0.15 (0.13, 0.17)	0.18 (0.16, 0.20)
Migration rates				
$ra \rightarrow rb$ (or $w \rightarrow z$)	0.557 (0.515, 0.603)	0.594 (0.549, 0.641)	0.531 (0.489, 0.574)	0.562 (0.516, 0.605)
$ac \rightarrow rb$ (or $x \rightarrow y$)	0.016 (0.003, 0.029)	0.009 (0.000, 0.020)	0.016 (0.004, 0.029)	0.021 (0.007, 0.036)

Note.– See Note in table S3.2. Node o is the root of the species tree including the outgroup. The estimates for the first half with the *D. insularis* outgroup are used to simulate data analysed in table S3.8.

Table S3.4: Posterior means and 95% HPD CIs of parameters under the final MSC-I and MSC-M models of Figure 3.1b obtained from BPP analyses of the *Drosophila* data set

	MSC-I		MSC-M	
Parameters	First half	Second half	First half	Second half
Population sizes (θ, $\times 10^{-2}$)				
θ_r	12.85 (11.95, 13.76)	14.47 (13.50, 15.46)	13.24 (12.22, 14.29)	14.86 (13.81, 15.92)
θ_a	5.19 (4.92, 5.46)	5.22 (4.98, 5.48)	4.62 (4.41, 4.82)	4.64 (4.44, 4.86)
θ_b	7.49 (7.03, 7.95)	8.16 (7.68, 8.65)	6.92 (6.51, 7.31)	7.74 (7.28, 8.19)
θ_c	5.28 (4.31, 6.30)	4.44 (3.64, 5.27)	5.16 (4.25, 6.09)	4.34 (3.60, 5.10)
θ_d	1.72 (1.62, 1.82)	1.77 (1.67, 1.88)	1.74 (1.64, 1.85)	1.79 (1.68, 1.89)
θ_e	3.78 (3.54, 4.02)	3.74 (3.50, 3.99)	3.79 (3.56, 4.03)	3.75 (3.51, 3.99)
θ_f	1.51 (1.41, 1.61)	1.48 (1.38, 1.58)	1.52 (1.42, 1.62)	1.49 (1.39, 1.59)
θ_g	1.09 (0.95, 1.23)	1.12 (0.97, 1.27)	1.09 (0.95, 1.23)	1.12 (0.97, 1.27)
θ_h	1.93 (1.78, 2.07)	1.88 (1.75, 2.02)	1.92 (1.78, 2.07)	1.88 (1.75, 2.02)
θ_i	1.46 (1.24, 1.70)	1.31 (1.07, 1.55)	1.46 (1.23, 1.70)	1.30 (1.06, 1.54)
θ_w	$= \theta_a$	$= \theta_a$	–	–
θ_z	$= \theta_b$	$= \theta_b$	–	–
θ_x	$= \theta_c$	$= \theta_c$	–	–
θ_y	$= \theta_b$	$= \theta_b$	–	–
Speciation/introgression times (τ, $\times 10^{-2}$)				
τ_r	7.28 (6.89, 7.74)	7.42 (7.10, 7.73)	7.48 (6.93, 8.17)	7.61 (7.05, 8.25)
τ_a	2.58 (2.54, 2.62)	2.58 (2.54, 2.62)	2.60 (2.57, 2.64)	2.61 (2.57, 2.65)
τ_b	2.21 (2.18, 2.24)	2.19 (2.15, 2.22)	2.23 (2.19, 2.26)	2.21 (2.17, 2.25)
τ_c	1.85 (1.77, 1.94)	1.94 (1.88, 2.01)	1.86 (1.78, 1.93)	1.95 (1.89, 2.01)
τ_d	1.03 (1.01, 1.05)	1.01 (0.99, 1.04)	1.03 (1.01, 1.05)	1.01 (0.99, 1.03)
τ_e	1.02 (0.99, 1.05)	1.03 (1.00, 1.06)	1.02 (0.99, 1.05)	1.03 (1.00, 1.06)
τ_f	0.69 (0.67, 0.72)	0.73 (0.70, 0.75)	0.69 (0.66, 0.71)	0.72 (0.70, 0.75)
τ_g	0.63 (0.61, 0.66)	0.62 (0.59, 0.65)	0.63 (0.61, 0.66)	0.62 (0.59, 0.65)
τ_h	0.41 (0.39, 0.43)	0.40 (0.38, 0.42)	0.41 (0.39, 0.43)	0.40 (0.38, 0.42)
τ_i	0.15 (0.12, 0.17)	0.18 (0.16, 0.20)	0.15 (0.12, 0.17)	0.18 (0.16, 0.20)
$\tau_w = \tau_z = \tau_{w \to z}$	3.81 (3.75, 3.87)	3.88 (3.82, 3.94)	–	–
$\tau_x = \tau_y = \tau_{x \to y}$	2.57 (2.53, 2.61)	2.57 (2.53, 2.61)	–	–
Gene-flow rate	Introgression probability (φ)		Migration rate (M)	
$ra \to rb$ (or $w \to z$)	0.728 (0.689, 0.769)	0.712 (0.681, 0.743)	0.568 (0.518, 0.615)	0.603 (0.555, 0.652)
$ac \to rb$ (or $x \to y$)	0.069 (0.055, 0.083)	0.069 (0.056, 0.082)	0.011 (0.000, 0.025)	0.009 (0.000, 0.020)

Table S3.5: Posterior means and 95% HPD CIs of introgression probabilities (φ), introgression times (τ) and Bayes factors in support of gene flow (B_{10}) in the BPP analysis of the *Drosophila* data under the MSC-I model with $w \to z$ and $x \to y$ introgressions, and bidirectional introgression between extant species X and Y

	First half (1389 loci)			Second half (1388 loci)		
Introgression	$\hat{\varphi}$	$\hat{\tau}$	B_{10}	$\hat{\varphi}$	$\hat{\tau}$	B_{10}
$X =$ *D. obscura* and $Y =$ *D. pseudoobscura*						
$w \to z$	0.7461 (0.6968, 0.7990)	0.0380 (0.0374, 0.0387)	∞	0.7029 (0.6738, 0.7313)	0.0387 (0.0381, 0.0392)	∞
$x \to y$	0.0669 (0.0522, 0.0818)	0.0258 (0.0253, 0.0262)	∞	0.0671 (0.0545, 0.0803)	0.0257 (0.0253, 0.0261)	∞
$X \to Y$	0.0011 (0.0000, 0.0034)	0.0007 (0.0000, 0.0014)	0.01	0.0009 (0.0000, 0.0028)	0.0009 (0.0000, 0.0017)	0.01
$Y \to X$	0.0014 (0.0000, 0.0036)		0.01	0.0008 (0.0000, 0.0024)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. pseudoobscura*						
$w \to z$	0.7345 (0.7005, 0.7683)	0.0380 (0.0374, 0.0386)	∞	0.7014 (0.6714, 0.7311)	0.0387 (0.0381, 0.0392)	∞
$x \to y$	0.0675 (0.0534, 0.0820)	0.0257 (0.0253, 0.0261)	∞	0.0672 (0.0547, 0.0803)	0.0257 (0.0253, 0.0262)	∞
$X \to Y$	0.0009 (0.0000, 0.0027)	0.0007 (0.0000, 0.0014)	0.01	0.0008 (0.0000, 0.0026)	0.0009 (0.0000, 0.0017)	0.01
$Y \to X$	0.0014 (0.0000, 0.0035)		0.01	0.0007 (0.0000, 0.0022)		0.01
$X =$ *D. obscura* and $Y =$ *D. persimilis*						
$w \to z$	0.7215 (0.6838, 0.7579)	0.0380 (0.0374, 0.0386)	∞	0.7000 (0.6592, 0.7331)	0.0386 (0.0380, 0.0392)	∞
$x \to y$	0.0659 (0.0520, 0.0804)	0.0257 (0.0253, 0.0261)	∞	0.0670 (0.0543, 0.0800)	0.0257 (0.0253, 0.0261)	∞
$X \to Y$	0.0011 (0.0000, 0.0033)	0.0008 (0.0000, 0.0014)	0.01	0.0011 (0.0000, 0.0034)	0.0009 (0.0000, 0.0017)	0.01
$Y \to X$	0.0015 (0.0000, 0.0037)		0.01	0.0008 (0.0000, 0.0024)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. persimilis*						
$w \to z$	0.7338 (0.6977, 0.7686)	0.0380 (0.0374, 0.0386)	∞	0.7064 (0.6776, 0.7353)	0.0386 (0.0381, 0.0392)	∞
$x \to y$	0.0673 (0.0532, 0.0822)	0.0257 (0.0253, 0.0261)	∞	0.0669 (0.0543, 0.0803)	0.0257 (0.0253, 0.0262)	∞
$X \to Y$	0.0009 (0.0000, 0.0029)	0.0008 (0.0000, 0.0014)	0.01	0.0009 (0.0000, 0.0027)	0.0009 (0.0000, 0.0017)	0.01
$Y \to X$	0.0014 (0.0000, 0.0035)		0.01	0.0007 (0.0000, 0.0023)		0.01
$X =$ *D. obscura* and $Y =$ *D. miranda*						
$w \to z$	0.7239 (0.6740, 0.7688)	0.0380 (0.0374, 0.0386)	∞	0.7023 (0.6720, 0.7315)	0.0386 (0.0380, 0.0392)	∞
$x \to y$	0.0662 (0.0524, 0.0810)	0.0257 (0.0253, 0.0261)	∞	0.0676 (0.0550, 0.0809)	0.0256 (0.0252, 0.0260)	∞
$X \to Y$	0.0010 (0.0000, 0.0031)	0.0023 (0.0001, 0.0041)	0.01	0.0010 (0.0000, 0.0030)	0.0020 (0.0000, 0.0039)	0.01
$Y \to X$	0.0014 (0.0000, 0.0037)		0.01	0.0008 (0.0000, 0.0024)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. miranda*						
$w \to z$	0.7198 (0.6858, 0.7546)	0.0379 (0.0373, 0.0385)	∞	0.7057 (0.6768, 0.7351)	0.0387 (0.0381, 0.0393)	∞
$x \to y$	0.0669 (0.0528, 0.0818)	0.0257 (0.0252, 0.0261)	∞	0.0675 (0.0547, 0.0806)	0.0257 (0.0253, 0.0262)	∞
$X \to Y$	0.0009 (0.0000, 0.0028)	0.0022 (0.0002, 0.0041)	0.01	0.0009 (0.0000, 0.0028)	0.0020 (0.0000, 0.0039)	0.01
$Y \to X$	0.0014 (0.0000, 0.0035)		0.01	0.0007 (0.0000, 0.0023)		0.01
$X =$ *D. obscura* and $Y =$ *D. lowei*						
$w \to z$	0.7134 (0.6762, 0.7523)	0.0380 (0.0374, 0.0386)	∞	0.7220 (0.6902, 0.7524)	0.0387 (0.0381, 0.0393)	∞
$x \to y$	0.0654 (0.0512, 0.0798)	0.0257 (0.0253, 0.0261)	∞	0.0679 (0.0550, 0.0811)	0.0258 (0.0254, 0.0262)	∞
$X \to Y$	0.0008 (0.0000, 0.0025)	0.0056 (0.0006, 0.0102)	0.01	0.0008 (0.0000, 0.0025)	0.0057 (0.0006, 0.0103)	0.01
$Y \to X$	0.0014 (0.0000, 0.0039)		0.01	0.0008 (0.0000, 0.0025)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. lowei*						
$w \to z$	0.7132 (0.6705, 0.7559)	0.0379 (0.0373, 0.0385)	∞	0.7178 (0.6868, 0.7473)	0.0387 (0.0381, 0.0393)	∞
$x \to y$	0.0662 (0.0522, 0.0811)	0.0257 (0.0253, 0.0261)	∞	0.0678 (0.0547, 0.0807)	0.0257 (0.0253, 0.0262)	∞
$X \to Y$	0.0007 (0.0000, 0.0023)	0.0048 (0.0000, 0.0094)	0.01	0.0008 (0.0000, 0.0024)	0.0051 (0.0000, 0.0097)	0.01
$Y \to X$	0.0014 (0.0000, 0.0036)		0.01	0.0008 (0.0000, 0.0024)		0.01

	First half (1389 loci)			Second half (1388 loci)		
Introgression	$\hat{\varphi}$	$\hat{\tau}$	B_{10}	$\hat{\varphi}$	$\hat{\tau}$	B_{10}
$X =$ *D. obscura* and $Y =$ *D. azteca*						
$w \to z$	0.7113 (0.6768, 0.7461)	0.0379 (0.0373, 0.0385)	∞	0.6967 (0.6681, 0.7249)	0.0387 (0.0381, 0.0392)	∞
$x \to y$	0.0661 (0.0521, 0.0807)	0.0257 (0.0253, 0.0261)	∞	0.0672 (0.0546, 0.0802)	0.0257 (0.0253, 0.0262)	∞
$X \to Y$	0.0008 (0.0000, 0.0025)	0.0058 (0.0005, 0.0103)	0.01	0.0007 (0.0000, 0.0024)	0.0051 (0.0003, 0.0100)	0.01
$Y \to X$	0.0010 (0.0000, 0.0030)		0.01	0.0008 (0.0000, 0.0024)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. azteca*						
$w \to z$	0.7279 (0.6892, 0.7691)	0.0380 (0.0374, 0.0386)	∞	0.7042 (0.6760, 0.7323)	0.0387 (0.0381, 0.0392)	∞
$x \to y$	0.0669 (0.0527, 0.0813)	0.0257 (0.0253, 0.0261)	∞	0.0671 (0.0543, 0.0799)	0.0257 (0.0253, 0.0261)	∞
$X \to Y$	0.0007 (0.0000, 0.0023)	0.0054 (0.0005, 0.0103)	0.01	0.0007 (0.0000, 0.0023)	0.0050 (0.0003, 0.0099)	0.01
$Y \to X$	0.0007 (0.0000, 0.0023)		0.01	0.0008 (0.0000, 0.0024)		0.01
$X =$ *D. obscura* and $Y =$ *D. athabasca*						
$w \to z$	0.7299 (0.6866, 0.7747)	0.0380 (0.0374, 0.0386)	∞	0.7257 (0.6960, 0.7545)	0.0387 (0.0381, 0.0393)	∞
$x \to y$	0.0669 (0.0527, 0.0815)	0.0257 (0.0253, 0.0261)	∞	0.0694 (0.0562, 0.0825)	0.0256 (0.0252, 0.0260)	∞
$X \to Y$	0.0008 (0.0000, 0.0025)	0.0033 (0.0002, 0.0063)	0.01	0.0008 (0.0000, 0.0024)	0.0031 (0.0000, 0.0059)	0.01
$Y \to X$	0.0008 (0.0000, 0.0026)		0.01	0.0008 (0.0000, 0.0026)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. athabasca*						
$w \to z$	0.7311 (0.6882, 0.7832)	0.0380 (0.0374, 0.0386)	∞	0.7018 (0.6735, 0.7318)	0.0387 (0.0381, 0.0392)	∞
$x \to y$	0.0669 (0.0525, 0.0814)	0.0257 (0.0253, 0.0261)	∞	0.0673 (0.0545, 0.0802)	0.0257 (0.0253, 0.0262)	∞
$X \to Y$	0.0008 (0.0000, 0.0024)	0.0032 (0.0001, 0.0061)	0.01	0.0008 (0.0000, 0.0025)	0.0027 (0.0000, 0.0058)	0.01
$Y \to X$	0.0007 (0.0000, 0.0023)		0.01	0.0010 (0.0000, 0.0029)		0.01
$X =$ *D. obscura* and $Y =$ *D. affinis*						
$w \to z$	0.7385 (0.7017, 0.7781)	0.0381 (0.0375, 0.0387)	∞	0.7020 (0.6736, 0.7300)	0.0386 (0.0381, 0.0392)	∞
$x \to y$	0.0675 (0.0529, 0.0821)	0.0257 (0.0253, 0.0262)	∞	0.0669 (0.0541, 0.0799)	0.0257 (0.0253, 0.0261)	∞
$X \to Y$	0.0009 (0.0000, 0.0027)	0.0001 (0.0000, 0.0002)	0.01	0.0008 (0.0000, 0.0026)	0.0001 (0.0000, 0.0003)	0.01
$Y \to X$	0.0008 (0.0000, 0.0026)		0.01	0.0008 (0.0000, 0.0024)		0.01
$X =$ *D. bifasciata* and $Y =$ *D. affinis*						
$w \to z$	0.7139 (0.6794, 0.7487)	0.0379 (0.0374, 0.0386)	∞	0.7026 (0.6750, 0.7295)	0.0386 (0.0381, 0.0392)	∞
$x \to y$	0.0665 (0.0523, 0.0812)	0.0257 (0.0253, 0.0261)	∞	0.0671 (0.0545, 0.0803)	0.0258 (0.0253, 0.0262)	∞
$X \to Y$	0.0008 (0.0000, 0.0025)	0.0000 (0.0000, 0.0001)	0.01	0.0008 (0.0000, 0.0025)	0.0002 (0.0000, 0.0004)	0.01
$Y \to X$	0.0007 (0.0000, 0.0022)		0.01	0.0008 (0.0000, 0.0026)		0.01
$X =$ *D. lowei*, $Y =$ *D. azteca*, $Z =$ *D. affinis*						
$w \to z$	0.7248 (0.6826, 0.7677)	0.0380 (0.0374, 0.0386)	∞	0.7061 (0.6767, 0.7348)	0.0388 (0.0382, 0.0394)	∞
$x \to y$	0.0671 (0.0530, 0.0817)	0.0257 (0.0253, 0.0260)	∞	0.0684 (0.0557, 0.0815)	0.0257 (0.0253, 0.0261)	∞
$X \to Y$	0.0028 (0.0002, 0.0060)	0.0055 (0.0020, 0.0081)	0.01	0.0009 (0.0000, 0.0028)	0.0073 (0.0070, 0.0075)	0.01
$Y \to X$	0.0008 (0.0000, 0.0024)		0.01	0.0019 (0.0000, 0.0046)		0.01
$X \to Z$	0.0022 (0.0000, 0.0054)	0.0026 (0.0003, 0.0058)	0.01	0.0010 (0.0000, 0.0032)	0.0008 (0.0000, 0.0035)	0.01
$Z \to X$	0.0007 (0.0000, 0.0023)		0.01	0.0007 (0.0000, 0.0023)		0.01

Note.– All tested gene flow events between extant species are rejected when the model already accounts for the $w \to z$ and $x \to y$ introgressions.

Table S3.6: Posterior means and 95% HPD CIs (in parentheses) obtained in BPP analyses of the real data under the JC and GTR mutation models

	JC		GTR	
Model	First half	Second half	First half	Second half
MSC-I				
$\varphi_{w\to z}$	0.728 (0.689, 0.769)	0.712 (0.681, 0.743)	0.723 (0.688, 0.758)	0.708 (0.675, 0.739)
$\varphi_{x\to y}$	0.069 (0.055, 0.083)	0.069 (0.056, 0.082)	0.067 (0.053, 0.081)	0.069 (0.056, 0.082)
MSC-M				
$M_{w\to z}$	0.568 (0.518, 0.615)	0.603 (0.555, 0.652)	0.568 (0.523, 0.614)	0.605 (0.555, 0.654)
$M_{x\to y}$	0.011 (0.000, 0.025)	0.009 (0.000, 0.020)	0.008 (0.000, 0.018)	0.007 (0.000, 0.016)

Note.– Estimates of τs and θs are shown in figure S3.2.

Table S3.7: Parameter estimates (posterior means and 95% HPD CIs) and Bayes factors (B_{10}) for testing gene flow in BPP analysis of triplet and quintet datasets informative for *D. lowei* ↔ *D. affinis* gene flow

	X = *D. pseudoobscura*		X = *D. persimilis*		X = *D. miranda*	
Parameter	Estimate	B_{10}	Estimate	B_{10}	Estimate	B_{10}
Triplet tree: ((*X*, *D. lowei*), *D. affinis*)						
$\hat{\varphi}_{p\to q}$ (*D. lowei* → *D. affinis*)	0.0477 (0.0326, 0.0630)	∞	0.0478 (0.0333, 0.0631)	∞	0.0420 (0.0307, 0.0537)	∞
$\hat{\varphi}_{q\to p}$ (*D. affinis* → *D. lowei*)	0.0089 (0.0000, 0.0186)	0.01	0.0046 (0.0000, 0.0130)	0.01	0.0010 (0.0000, 0.0029)	0.01
τ_b	0.0198 (0.0193, 0.0202)		0.0192 (0.0187, 0.0196)		0.0199 (0.0195, 0.0203)	
τ_e	0.0085 (0.0083, 0.0088)		0.0086 (0.0083, 0.0088)		0.0084 (0.0082, 0.0086)	
$\tau_p = \tau_q$	0.0082 (0.0074, 0.0087)		0.0081 (0.0072, 0.0087)		0.0081 (0.0074, 0.0086)	
$\theta_{D.\ lowei}$	0.0136 (0.0013, 0.0293)		0.0134 (0.0012, 0.0292)		0.0150 (0.0016, 0.0320)	
$\theta_{D.\ affinis}$	0.0147 (0.0018, 0.0298)		0.0143 (0.0005, 0.0307)		0.0109 (0.0002, 0.0260)	
θ_b	0.0609 (0.0587, 0.0631)		0.0662 (0.0639, 0.0686)		0.0458 (0.0439, 0.0476)	
θ_e	0.0221 (0.0206, 0.0236)		0.0229 (0.0213, 0.0245)		0.0130 (0.0122, 0.0137)	
Quintet tree: (((*X*, *D. lowei*), *D. affinis*), (*D. obscura*, *D. guanche*))						
$\hat{\varphi}_{p\to q}$ (*D. lowei* → *D. affinis*)	0.0148 (0.0069, 0.0232)	0.08	0.0231 (0.0132, 0.0332)	6.21	0.0169 (0.0093, 0.0246)	0.35
$\hat{\varphi}_{q\to p}$ (*D. affinis* → *D. lowei*)	0.0010 (0.0000, 0.0030)	0.01	0.0009 (0.0000, 0.0026)	0.01	0.0008 (0.0000, 0.0024)	0.01
τ_r	0.0298 (0.0295, 0.0301)		0.0295 (0.0292, 0.0299)		0.0301 (0.0298, 0.0304)	
τ_a	0.0208 (0.0202, 0.0214)		0.0208 (0.0202, 0.0215)		0.0214 (0.0207, 0.0221)	
τ_b	0.0207 (0.0203, 0.0210)		0.0204 (0.0199, 0.0208)		0.0206 (0.0202, 0.0209)	
τ_e	0.0093 (0.0090, 0.0095)		0.0093 (0.0090, 0.0095)		0.0090 (0.0088, 0.0092)	
$\tau_p = \tau_q$	0.0086 (0.0072, 0.0095)		0.0089 (0.0079, 0.0095)		0.0085 (0.0074, 0.0092)	
$\theta_{D.\ lowei}$	0.0137 (0.0012, 0.0296)		0.0145 (0.0015, 0.0309)		0.0144 (0.0014, 0.0308)	
$\theta_{D.\ affinis}$	0.0114 (0.0003, 0.0271)		0.0115 (0.0003, 0.0273)		0.0112 (0.0003, 0.0266)	
θ_r	0.0762 (0.0742, 0.0782)		0.0785 (0.0764, 0.0805)		0.0662 (0.0644, 0.0680)	
θ_a	0.0743 (0.0646, 0.0843)		0.0702 (0.0611, 0.0796)		0.0838 (0.0714, 0.0966)	
θ_b	0.0461 (0.0428, 0.0495)		0.0500 (0.0462, 0.0537)		0.0307 (0.0285, 0.0328)	
θ_e	0.0216 (0.0202, 0.0229)		0.0229 (0.0214, 0.0244)		0.0128 (0.0121, 0.0136)	
Quintet tree: (((*X*, *D. lowei*), *D. affinis*), (*D. obscura*, *D. guanche*)), with $w \to z$ and $x \to y$ introgressions						
$\hat{\varphi}_{w\to z}$	0.9066 (0.8931, 0.9196)	∞	0.8750 (0.8600, 0.8901)	∞	0.9513 (0.9405, 0.9621)	∞
$\hat{\varphi}_{x\to y}$	0.0890 (0.0754, 0.1031)	∞	0.0851 (0.0717, 0.0981)	∞	0.0913 (0.0764, 0.1063)	∞
$\hat{\varphi}_{p\to q}$ (*D. lowei* → *D. affinis*)	0.0413 (0.0273, 0.0555)	∞	0.0571 (0.0418, 0.0723)	∞	0.0496 (0.0365, 0.0629)	∞
$\hat{\varphi}_{q\to p}$ (*D. affinis* → *D. lowei*)	0.0009 (0.0000, 0.0027)	0.01	0.0010 (0.0000, 0.0031)	0.01	0.0007 (0.0000, 0.0023)	0.01
τ_r	0.0902 (0.0852, 0.0949)		0.0818 (0.0795, 0.0841)		0.0872 (0.0837, 0.0906)	
τ_a	0.0225 (0.0222, 0.0229)		0.0225 (0.0221, 0.0229)		0.0234 (0.0230, 0.0238)	
τ_b	0.0205 (0.0200, 0.0209)		0.0205 (0.0201, 0.0210)		0.0210 (0.0205, 0.0214)	
τ_e	0.0090 (0.0088, 0.0093)		0.0090 (0.0087, 0.0092)		0.0088 (0.0086, 0.0090)	
$\tau_w = \tau_z$	0.0356 (0.0351, 0.0360)		0.0353 (0.0348, 0.0357)		0.0352 (0.0348, 0.0357)	
$\tau_x = \tau_y$	0.0225 (0.0221, 0.0228)		0.0224 (0.0220, 0.0228)		0.0233 (0.0229, 0.0237)	
$\tau_p = \tau_q$	0.0088 (0.0084, 0.0093)		0.0088 (0.0085, 0.0092)		0.0087 (0.0083, 0.0090)	
$\theta_{D.\ obscura}$	0.0018 (0.0001, 0.0042)		0.0013 (0.0000, 0.0029)		0.0011 (0.0001, 0.0025)	
$\theta_{D.\ lowei}$	0.0149 (0.0013, 0.0316)		0.0148 (0.0015, 0.0317)		0.0149 (0.0017, 0.0319)	
$\theta_{D.\ affinis}$	0.0116 (0.0003, 0.0273)		0.0118 (0.0003, 0.0277)		0.0112 (0.0002, 0.0266)	
θ_r	0.1228 (0.1130, 0.1333)		0.1130 (0.1049, 0.1211)		0.1183 (0.1071, 0.1299)	
θ_a	0.0493 (0.0475, 0.0512)		0.0483 (0.0464, 0.0501)		0.0491 (0.0473, 0.0508)	
θ_b	0.0476 (0.0447, 0.0505)		0.0510 (0.0479, 0.0543)		0.0305 (0.0284, 0.0325)	
θ_e	0.0221 (0.0207, 0.0235)		0.0237 (0.0222, 0.0253)		0.0134 (0.0126, 0.0142)	
θ_w	$= \theta_a$		$= \theta_a$		$= \theta_a$	
θ_z	$= \theta_b$		$= \theta_b$		$= \theta_b$	
θ_x	$= \theta_{D.\ obscura}$		$= \theta_{D.\ obscura}$		$= \theta_{D.\ obscura}$	
θ_y	$= \theta_b$		$= \theta_b$		$= \theta_b$	
θ_p	$= \theta_{D.\ lowei}$		$= \theta_{D.\ lowei}$		$= \theta_{D.\ lowei}$	
θ_q	$= \theta_{D.\ affinis}$		$= \theta_{D.\ affinis}$		$= \theta_{D.\ affinis}$	

Note.– Nodes r, a, b, e, p, q refer to the species tree in figure 3.1a. The triplet species tree is ((*X*, *D. lowei*), *D. affinis*), where *X* is *D. pseudoobscura*, *D. persimilis*, or *D. miranda*, with species divergences at nodes *b* and *e* and with bidirectional introgression at nodes p, q (Fig. 3.1a). The quintet datasets include two outgroup species *D. guanche* and *D. obscura*, with an additional internal node *a* (Fig. 3.1a). The quartet introgression model assumes the $w \to z$ and $x \to y$ introgressions (as in Fig. 3.1b). The number of loci in the triplet datasets is 2672, 2646, and 2672 for $X =$ *D. pseudoobscura*, *D. persimilis*, and *D. miranda*, respectively, and the corresponding numbers for the quintet datasets are 2759, 2757, and 2758.

Table S3.8: Posterior means and 95% HPD CIs (in parentheses) from BPP analysis of datasets simulated under the MSC-I and MSC-M models of figure 3.1b with the *D. insularis* outgroup

	MSC-I			MSC-M		
Parameters	**Truth**	Estimates, first half	Estimates, second half	**Truth**	Estimates, first half	Estimates, second half
Population sizes (θ, $\times 10^{-2}$)						
θ_o	**17.95**	17.61 (16.79, 18.44)	17.53 (16.72, 18.36)	**18.06**	17.85 (17.02, 18.69)	17.28 (16.44, 18.06)
θ_r	**11.39**	10.94 (9.18, 12.69)	11.30 (9.88, 12.73)	**6.99**	1.83 (0.17, 4.03)	1.73 (0.15, 3.78)
θ_a	**5.14**	5.13 (4.86, 5.40)	5.17 (4.89, 5.46)	**4.81**	4.65 (4.42, 4.87)	4.88 (4.65, 5.11)
θ_b	**7.20**	7.15 (6.62, 7.68)	6.64 (6.16, 7.13)	**6.83**	6.71 (6.27, 7.15)	6.49 (6.05, 6.96)
θ_c	**4.50**	3.53 (2.63, 4.43)	3.50 (2.44, 4.60)	**4.55**	4.19 (3.04, 5.40)	5.19 (3.90, 6.53)
θ_d	**1.72**	1.84 (1.69, 2.00)	1.72 (1.58, 1.86)	**1.73**	1.79 (1.64, 1.94)	1.82 (1.67, 1.98)
θ_e	**3.72**	3.82 (3.51, 4.14)	4.02 (3.70, 4.33)	**3.72**	3.87 (3.56, 4.18)	3.68 (3.39, 3.98)
θ_f	**1.57**	1.57 (1.44, 1.69)	1.55 (1.42, 1.67)	**1.58**	1.48 (1.36, 1.60)	1.61 (1.48, 1.74)
θ_g	**1.11**	0.96 (0.66, 1.26)	1.27 (0.97, 1.58)	**1.10**	1.08 (0.77, 1.40)	1.03 (0.75, 1.32)
θ_h	**1.92**	1.89 (1.66, 2.13)	1.70 (1.48, 1.92)	**1.92**	1.95 (1.69, 2.21)	2.04 (1.79, 2.29)
θ_i	**1.46**	1.69 (1.16, 2.22)	1.75 (1.26, 2.24)	**1.46**	1.68 (1.26, 2.13)	1.13 (0.70, 1.61)
θ_w	$= \theta_a$			**–**	–	–
θ_z	$= \theta_b$			**–**	–	–
θ_x	$= \theta_c$			**–**	–	–
θ_y	$= \theta_b$			**–**	–	–
Speciation/introgression times (τ, $\times 10^{-2}$)						
τ_o	**9.54**	9.57 (9.39, 9.75)	9.67 (9.49, 9.84)	**9.51**	9.37 (9.20, 9.53)	9.63 (9.47, 9.78)
τ_r	**7.53**	7.78 (7.40, 8.17)	7.37 (6.97, 7.77)	**8.77**	9.32 (9.09, 9.54)	9.60 (9.41, 9.78)
τ_a	**2.75**	2.78 (2.72, 2.84)	2.73 (2.68, 2.79)	**2.74**	2.69 (2.63, 2.74)	2.73 (2.68, 2.78)
τ_b	**2.18**	2.20 (2.15, 2.26)	2.24 (2.19, 2.30)	**2.20**	2.19 (2.14, 2.24)	2.24 (2.18, 2.29)
τ_c	**1.96**	2.07 (1.95, 2.20)	2.11 (1.97, 2.26)	**1.95**	1.98 (1.84, 2.12)	1.90 (1.75, 2.04)
τ_d	**1.03**	1.00 (0.96, 1.05)	1.02 (0.98, 1.06)	**1.02**	1.01 (0.97, 1.06)	1.03 (0.98, 1.07)
τ_e	**1.01**	1.01 (0.96, 1.06)	0.97 (0.92, 1.02)	**1.01**	0.98 (0.93, 1.03)	1.01 (0.96, 1.05)
τ_f	**0.75**	0.76 (0.71, 0.81)	0.77 (0.72, 0.81)	**0.74**	0.77 (0.72, 0.81)	0.72 (0.67, 0.77)
τ_g	**0.63**	0.66 (0.59, 0.73)	0.60 (0.53, 0.66)	**0.63**	0.64 (0.58, 0.71)	0.66 (0.59, 0.72)
τ_h	**0.41**	0.41 (0.37, 0.45)	0.43 (0.39, 0.46)	**0.41**	0.42 (0.38, 0.46)	0.39 (0.35, 0.43)
τ_i	**0.15**	0.12 (0.06, 0.17)	0.11 (0.06, 0.16)	**0.15**	0.10 (0.05, 0.15)	0.17 (0.12, 0.23)
$\tau_{w\to z}$	**4.00**	4.04 (3.96, 4.12)	4.01 (3.93, 4.09)	**–**	–	–
$\tau_{x\to y}$	**2.73**	2.75 (2.68, 2.81)	2.69 (2.60, 2.77)	**–**	–	–
Gene-flow rate		Introgression probability (φ)			Migration rate (M)	
$ra \to rb$ (or $w \to z$)	**0.708**	0.716 (0.686, 0.747)	0.712 (0.680, 0.744)	**0.557**	0.500 (0.457, 0.544)	0.524 (0.472, 0.571)
$ac \to rb$ (or $x \to y$)	**0.098**	0.119 (0.094, 0.144)	0.090 (0.068, 0.112)	**0.016**	0.023 (0.002, 0.048)	0.033 (0.002, 0.068)

Note.– True parameter values for the MSC-I model with $w \to z$ and $x \to y$ introgressions are from table S3.2 (first half with the *D. insularis* outgroup), and the parameter values for the MSC-M model are from table S3.3. Node o is the root of the species tree including the outgroup.

Table S3.9: Posterior means and 95% HPD CIs (in parentheses) for the rate of gene flow (φ in MSC-I or M in MSC-M) obtained in BPP analyses of the simulated data using the 0.1x, 1x and 10x priors

		Priors on τ and θ		
Model	**Truth**	0.1x	1x	10x
MSC-I				
$\varphi_{w\to z}$	**0.708**	0.761 (0.734, 0.788)	0.716 (0.686, 0.747)	0.719 (0.688, 0.750)
$\varphi_{x\to y}$	**0.098**	0.123 (0.099, 0.147)	0.119 (0.094, 0.144)	0.119 (0.094, 0.143)
MSC-M				
$M_{w\to z}$	**0.557**	0.499 (0.458, 0.545)	0.500 (0.457, 0.544)	0.509 (0.465, 0.554)
$M_{x\to y}$	**0.016**	0.024 (0.001, 0.051)	0.023 (0.002, 0.048)	0.024 (0.001, 0.051)

Note.– Estimates of τs and θs are shown in figure S3.3. See legend to figure S3.3.

Chapter 4

Unravelling the Migration History between Chimpanzees and Bonobos

Bonobos (*Pan paniscus*) and chimpanzees (*Pan troglodytes*) are great apes found in western and central Africa, and they diverged from humans about 5-10 Mya (de Manuel *et al.*, 2016; Prado-Martinez *et al.*, 2013). The current taxonomy of the genus *Pan* recognises bonobos as one species, while chimpanzees are divided into four subspecies (Caswell *et al.*, 2008; Prado-Martinez *et al.*, 2013). The two species likely diverged through allopatric or sympatric speciation (Hey, 2010a; Osada and Wu, 2005) about 0.9 – 2 Mya ago. They are currently separated by the Congo river (fig. 4.1a) (de Manuel *et al.*, 2016; Kuhlwilm *et al.*, 2016b; Lobon *et al.*, 2016; Prado-Martinez *et al.*, 2013). Western chimpanzees, *P. t. verus*, occur in the most western part of the species geographic range, from Senegal on the west to Ghana on the east. The other three subspecies are separated from Western chimpanzees by the Dahomey gap. From west to east, Nigeria-Cameroon chimpanzees (*P. t. ellioti*) are separated from Central chimpanzees (*P. t. troglodytes*) by the Sanaga river, and Eastern chimpanzees (*P. t. schweinfurthii*) are separated from Central chimpanzees by the Ubangi river.

Genetic data have suggested that the four chimpanzee subspecies form two clades, which diverged about 400-600 Kya ago. Within *Pan troglodytes*, Western and Nigeria-Cameroon chimpanzees form one clade, with the estimated split time of about 250-500 Kya, while Central and Eastern chimpanzees form another clade, with the estimated split time of about 90-250 Kya (Becquet *et al.*, 2007; de Manuel *et al.*, 2016; Hey, 2010a; Kuhlwilm *et al.*, 2016b; Prado-Martinez *et al.*, 2013; Won and Hey, 2005). However, phylogenetic reconstruction may be affected by incomplete lineage sorting and gene flow between species/subspecies (Jiao *et al.*, 2021; Rannala *et al.*, 2020; Xu and Yang, 2016). Ignoring gene flow between populations may lead to serious underestimation of the divergence time (Ji *et al.*, 2023).

Analyses of genetic data under models of isolation with migration (IM) have found signals of gene flow between the chimpanzee subspecies, but there is little consensus regarding the subspecies

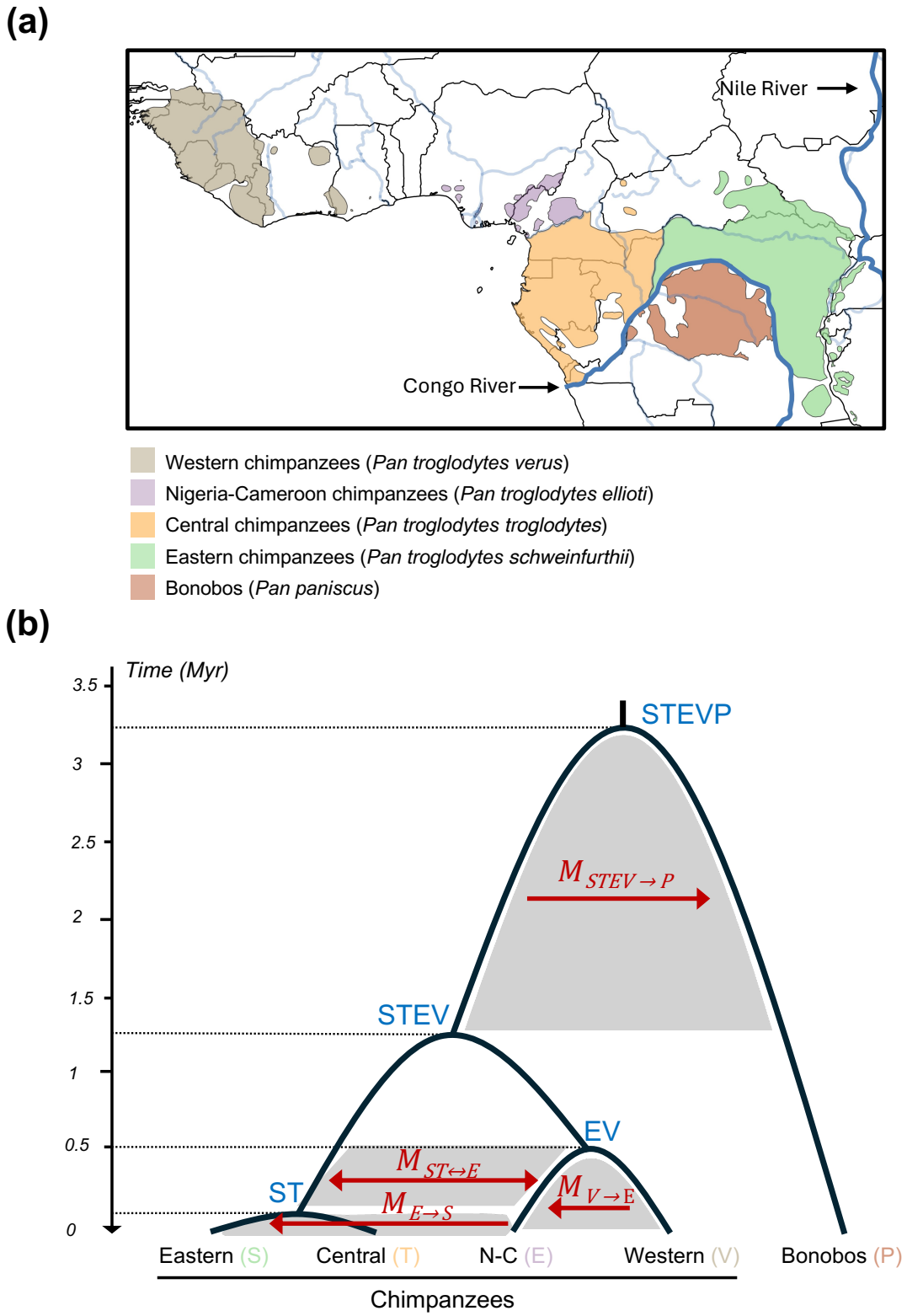

Figure 4.1: **(a)** Geographical distribution and **(b)** proposed migration events for bonobos (*Pan paniscus*) and four subspecies of chimpanzees (*Pan troglodytes*): western chimpanzees (*P. t. verus*), Nigeria-Cameroon chimpanzees (*P. t. ellioti*), central chimpanzees (*P. t. troglodytes*), and eastern chimpanzees (*P. t. schweinfurthii*). The time spans of migration events are indicated by grey shading on phylogeny. Time estimates τ are genome averages across noncoding blocks in BPP, which are converted into absolute times assuming mutation rate $\mu = 1.2 \times 10^{-8}$ per site per generation and a generation time of 25 years.

involved (see Figure 1 in Brand *et al.*, 2022 for a summary) (Becquet *et al.*, 2007; de Manuel *et al.*, 2016; Hey, 2010a; Hey *et al.*, 2018; Prado-Martinez *et al.*, 2013; Wegmann and Excoffier, 2010). Some gene-flow events are hard to reconcile with the current biogeography of the *Pan* genus (e.g., the introgression of 21% from Western to Eastern chimpanzees suggested by Brand *et al.*, 2022), although the authors argue that the current and historic ranges of these subspecies may be different. The geographic barriers (the Dahomey gap, the Sanaga and the Ubangi rivers) may be permeable in the past. The discharge of these rivers likely experienced large variations along the history of Congo rivers formed 1-2 Ma ago so that gene flow between bonobos and chimpanzees as well as geographically proximate chimpanzee subspecies may be possible during dry seasons (Brand *et al.*, 2022). There were large fluctuations in chimpanzee's ranges and substantial forest expansion around Dahomey Gap that potentially allowed for contact between the subspecies (McBrearty and Jablonski, 2005).

Early analyses of autosomal loci, incorporating data from three subspecies, have already detected rampant signals of migration across populations (Caswell *et al.*, 2008; Hey, 2010a; Wegmann and Excoffier, 2010). Models of gene flow proposed in these analyses present limited agreement, partly due to the lack of the Nigeria–Cameroon lineage. Recent analyses have used all five populations. For instance, de Manuel *et al.* (2016) identified multiple migration events between the two species, and among the chimpanzee subspecies, with the ancestral population of central and eastern chimpanzees being the recipient of bonobo alleles. Limited by computation, full likelihood methods have been applied to small datasets (Hey *et al.*, 2018). Estimation of gene flow with recently generated sequencing data using full likelihood methods may be essential to reliably resolve the history of population divergence and migration and introgression/hybridization in the *Pan* genus.

Brand *et al.* (2022) used a recently developed method called LEGOFIT (Rogers, 2019) to compare different models of gene flow between the bonobo and the chimpanzee subspecies. The method makes use of genome-wide site pattern frequencies, like the D-STATISTIC or HYDE. This suffers from multiple problems, such as failing to use samples from the same population and (i.e., ignoring within-population variation), ignoring variation in genealogical histories across the genome, etc. The method is not expected to have the capability of distinguishing among the different models considered

by Brand *et al.* (2022). Kuhlwilm *et al.* (2019) suggested that introgression into the bonobo lineage from a putative ancient great ape lineage, contributing up to 4.8% of the genome. This was done by fitting models to data of the (joint) site frequency spectrum (SFS) using FASTSIMCOAL. However, more proper testing in the likelihood framework using the IMA3 program found no evidence for existence of this ghost lineage Hey *et al.* (2018).

Here, we processed whole-genome sequencing data from de Manuel *et al.* (2016) for bonobos and chimpanzees to compile alignments of unphased diploid genomic sequences and used them to infer gene flow between species and subspecies. We compiled three sets of data, for exons (coding), noncoding regions, and conserved noncoding elements (CNEs). We use the full-likelihood method implemented in the BPP program (Flouri *et al.*, 2018). This includes an efficient Bayesian implementation of the MSC-M (or IM) model, which has been applied to genome-scale datasets with over 10,000 loci (Flouri *et al.*, 2023). We attempt to explain why different analyses of gene flow in bonobos and chimpanzees have produced highly incompatible results.

4.1 Materials and Methods

4.1.1 Genomic sequencing data and variant calling

We processed sequencing reads for chimpanzees and bonobos published by Prado-Martinez *et al.* (2013) and de Manuel *et al.* (2016) to compile multi-locus datasets comprised of aligned unphased diploid sequences. Prado-Martinez *et al.* (2013) sequenced 25 chimpanzee and 10 bonobo genomes, while de Manuel *et al.* (2016) sequenced 36 more chimpanzees. We used all samples except western chimpanzee sample Ptv-9730_Donald from de Manuel *et al.* (2016), which was identified as a central-western hybrid. There were thus 10 bonobo (*Pan paniscus*) samples and 60 chimpanzee (*Pan troglodytes*) samples, including 12 western chimpanzees (*P.t. verus*), 10 Nigeria-Cameroon chimpanzees (*P.t. ellioti*), 18 central chimpanzees (*P.t. troglodytes*), and 20 eastern chimpanzees (*P.t. schweinfurthii*). Most samples were sequenced to an average depth of at least 25x. The bonobo and the four chimpanzee subspecies are referred to below as five populations. We also retrieved two African human genomes (HG02615 and HG02975) from the 1000-Genomes Project (Byrska-Bishop

et al., 2022), and used them as outgroups.

Reads were mapped to the human reference genome hg19 for variant calling using BWA v0.7.7 (Li and Durbin, 2009). Genotypes were called using Genome Analysis Toolkit (GATK) v4.2.5.0 following GATK best practices (Poplin *et al.*, 2017). Variants were first called for each sample separately using GATK *HaplotypeCaller* resulting in a GVCF per sample. GVCFs were consolidated into a VCF using GATK *GenomicsDBImport* and joint genotype calling was performed using GATK *GenotypeGVCFs*. Indels were masked as missing data (alignment gaps, '-'). Variants were filtered using the `bcftools` v1.13 *filter* module (Li, 2011). We included multiallelic as well as biallelic SNPs. For each SNP, we required a minimum genotype quality (GQ) score of 20 and the read depth (DP) to be in the interval [20, 3*meanDP]. For the Y chromosome, the interval was [10, 3*meanDP]. All regions on the Y were non-recombining. Sites that did not meet those criteria were masked as missing data.

4.1.2 Selection of genomic regions and multilocus datasets

We compiled short genomic segments that are far apart in the genome and refer to them as loci. To ensure data quality, we used the coverage for each site in each sample. Three separate sets of data were generated, for coding, noncoding, and CNEs (conserved noncoding elements), respectively. Coding regions included exons from NCBI reference sequence database (RefSeq). CNEs are conserved intronic or intergenic regions, identified using PhastCons conservation scores calculated from genome alignments for 10 primate species (the primate subtrack) (Pollard *et al.*, 2010). CNE data were compiled first, before non-CNE noncoding loci, so that there is no overlap of loci between the two datasets.

For each individual, the per-site coverage was calculated using `samtools` v1.13 *depth* module (Danecek *et al.*, 2021). We then selected sites at which at least 10 individuals from at least four of the five species or subspecies had at least 8x coverage. Such high-coverage sites, if they are adjacent, were merged into regions. Regions with simple repeats, recent segmental duplications, CpG islands, and transposable elements as well as regions not showing conserved synteny in the human-chimpanzee alignment, based on annotations from the UCSC Genome Browser, were removed.

We used a coding/CNE region if it spanned over 100 sites or a noncoding region with over 500

sites. Each region was required to be at least 5kb away from the previous region. Long regions with more than 2kb were trimmed to 2kb by removing an equal number of nucleotides from both sides. An individual is said to be of high coverage for each region if $\geq 95\%$ of sites had a minimum coverage of 8x. A region is selected if at least 10 high-coverage individuals from at least four of the five populations were present, and at a selected region, low-coverage individuals are discarded. An unphased diploid sequence was constructed for each high-coverage individual in each region using the *consensus* module in `bcftools`, with heterozygotes represented using IUPAC ambiguity codes (e.g., Y for a T/C heterozygote). For chromosome X, we kept only female high-coverage individuals at each locus, as the male X chromosome tends to have low coverage and quality. For chromosome Y, we only considered male individuals and masked heterozygous genotype calls. If there exist a few sites with low coverage in a high-coverage individual, they were masked as missing data. Note that the number of sequences might be variable between loci, depending on the number of high-coverage individuals. Loci with $> 50\%$ missing data (across all sites and individuals at the locus) were excluded.

The number of loci and basic information for the three datasets (coding, noncoding and CNE) are shown in table S4.1. There are over 50,000 loci in each dataset, with more than 500 loci on each chromosome. Note that we generated alignments of unphased diploid sequences, rather than artificial haploid sequences with heterozygote phase resolved at random (Huang *et al.*, 2022b). When the data are analysed using BPP, the likelihood calculation averages over all possible phase resolutions at heterozygote sites using an analytical integration algorithm (Flouri *et al.*, 2018; Gronau *et al.*, 2011). At 556 coding, 13,065 noncoding, and 325 CNE loci, the number of site patterns in the A3 alignment (which includes all possible site patterns generated from all possible heterozygote phase resolutions) was > 3000. To reduce the computation load, we reduced the number of sequences at those loci to have two random samples per population.

4.1.3 Species trees across the genome

We inferred species trees of the five populations using Bayesian inference under the MSC model without gene flow implemented in BPP (Yang, 2015). For each of the three sets of data (coding,

noncoding and CNE), we formed 100-loci blocks in the order of occurrence in the human reference genome to infer the species tree under the MSC model with no gene flow. We assigned gamma priors to both population sizes and divergence times in the MSC model: $\theta \sim \text{G}(2, 2000)$ with mean 0.001, $\tau_0 \sim \text{G}(2, 400)$ with mean 0.005 for the age of the root (the human-chimpanzee divergence). We conducted 3 replicate MCMC runs, using different starting species trees. We used a burn-in of 20,000 iterations, and then took 2×10^5 samples, sampling every 2 iterations. Convergence was confirmed by ensuring that the MAP trees were identical in at least two runs, and its posterior probability differed by no more than 0.3. If those criteria were not met, the runs were repeated until convergence was achieved. MCMC samples from successful runs were combined to generate the posterior summaries.

4.1.4 Construction of a migration model

Our species tree analysis of the autosomes and the mitochondrial data suggested that the most likely population phylogeny was $(((S,T),(E,V)),P)$. We then took a two-step approach to add gene-flow events onto the population phylogeny to construct a model of gene flow for the bonobo and the four chimpanzee subspecies.

In the first step, we used the maximum likelihood program 3S (Dalquen *et al.*, 2017; Zhu and Yang, 2012) as well as BPP to analyse datasets of population triplets to infer gene flow. 3S implements likelihood ratio tests (LRTs) to test for gene flow among three closely related species/populations. It is computationally efficient and feasible with large datasets, but is limited to three populations. The method is used as an exploratory tool to assess the prevalence of gene flow between species and subspecies. We used bonobo (P) as the outgroup and chose two other populations from four chimpanzee subspecies, resulting in six triplets: STP, SEP, SVP, TEP, TVP, and EVP. Let the three populations be S_1, S_2, S_3, with the assumed phylogeny $((S_1, S_2), S_3)$. Each locus consists of three sequences. The sample configuration at each locus is selected at random, with probability 40% for 123, and probability 10% for each of the 6 configurations: 112, 122, 113, 133, 223, and 233, where '123' means one sequence for each population, '112' means two sequences from S_1 and one from S_2, and so on.

For each triplet, coding and CNE loci on the same chromosome were grouped into one dataset,

resulting in 23 datasets (for 22 autosomes and the X chromosome). Noncoding loci were on average five times longer (table S4.1) and contained more variable sites than exons and CNE loci. Thus, we sampled 4000 noncoding loci at random on each chromosome if there are more than 4000 noncoding loci, which similarly led to 23 datasets, each with at most 4000 loci.

Two models were fitted to each triplet dataset using 3s: the null MSC model with no gene flow (M0) and the MSC-migration model (MSC-M or M2). In M2, gene flow is specified between two chimpanzee subspecies and between the chimpanzee ancestor and the bonobo in both directions. For instance, for the triplet STP, the model assumes migration between populations S and T, and between their common ancestor ST and species P, with four migration rates ($M = Nm$). For each model, we performed 20 replicate runs, and the results corresponding to the highest log likelihood were used. Models M0 and M2 were compared using a likelihood ratio test. Migration rates were considered significant if they passed the LRT (for comparing M0 against M2) at the 1% level.

With the above locus sampling plan, the resulting dataset includes loci with configurations involving two sequences from one population, such as 122, 113, 133, 223, and 233. The inclusion of these configurations improves the power of the LRT of gene flow and is essential for accurate estimation of migration rates and population sizes (Dalquen *et al.*, 2017). This ensures that all parameters are identifiable in both models M0 and M2, with M2 containing more parameters (e.g., migration rates). As a result, the null distribution is known to be the 50:50 mixture of 0 and χ^2_k, where k is the number of migration rates in M2 (Dalquen *et al.*, 2017). Here, the null distribution is the mixture of 0 and χ^2_4, with 1% critical value to be 11.14. Note that the LRT in the context only suggests whether the migration model is favoured by the data over the null model without gene flow, while it has no power to identify which specific migration event has rate $M > 0$.

The Bayesian program BPP was also used to analyse triplet data under the MSC-M model, which allows the use of more than three sequences per locus. To reduce the computational cost, we selected five 100-loci blocks on each chromosome at random, with 23 datasets of noncoding, CNE and coding loci, each of 500 loci, for the 23 chromosomes. Data were prepared for the 6 triplets as above, including all sequences for the three populations. The triplet datasets were then analysed using BPP to fit the MSC-M model with the same four migration rates. The same gamma priors on θ and τ as in the

A01 analysis above were used, while the migration rate was assigned the gamma prior $M \sim \text{G}(1, 10)$ with mean 0.1. Each analysis was conducted twice, using a burn-in of 10^5 iterations and collecting 10^6 samples, sampling every 2 iterations. Convergence was confirmed by examining the consistency between runs. The significance of the migration rates was assessed by applying the Bayesian test of gene flow using the Savage-Dickey density ratio (Ji *et al.*, 2023), which can be used to compare nested models with (H_1) and without migration (H_0) based on an MCMC sample obtained under the model of gene flow (H_1). The Bayes factor B_{10} is calculated for each migration event in the model to determine whether the migration rate (M) is statistically excluded by a null interval $(0, \varepsilon)$. We used $\varepsilon = 0.01$ and confirmed that use of $\varepsilon = 0.001$ gave similar results. We used a cut-off of 100 for the Bayes factor ($B_{10} \geq 100$).

The triplet analyses suggested significant evidence for gene flow in many triplets (fig. S4.3). Note that the same gene-flow event involving ancestral lineages on the full population phylogeny may show up as significant evidence in many triplets, and also that multiple gene-flow events on the full phylogeny may be lumped into one event in the triplet analysis. We used the following criteria to integrate the results of the triplet analyses to formulate an MSC-M model for the full population phylogeny of five populations. Only events that were significant in both 3S and BPP tests on ≥ 5 chromosomes were retained in the model. If migration was found between species S_3 and both S_1 and S_2, we assume migration between S_3 and the common ancestor S_{12}. Migration events identified between E and ST can reconcile the two major species trees in the genome identified in the analysis of the 100-loci blocks (fig. 4.2).

The MSC-M model retaining migration events that passed those filters was then assumed to estimate species split times and migration rates using BPP.

4.1.5 Estimation of migration rates and species divergence times

Given the population phylogeny and the gene-flow events of figure 4.1b, we ran BPP to estimate the parameters: migration rates (M), species divergence times (τ) and population sizes (θ). The analysis was also performed based on data blocks, each of 200 loci. The blocks are twice as large as those used in species tree inference to achieve reasonable parameter estimates that are driven more by

data information than priors, while each block remains small still to reflect the local migration rate of that region. We assigned priors $\theta \sim$ G(2,2000), with mean 0.001, $\tau_0 \sim$ G(2,400), with mean 0.005, and $M \sim$ G(1,10) with mean 0.1. The MCMC was run for 10^6 iterations, sampling every 2 iterations, after a burn-in of 10^5 iterations.

4.2 Results

4.2.1 Fluctuation of genealogical relationships across the genome

We inferred the species tree using blocks of 100 loci under the MSC model with no gene flow. This is the A01 analysis of Yang (2015), and it accounts for ancestral polymorphism and incomplete lineage sorting but ignores gene flow. Note that chimpanzees and bonobos possess 24 pairs of chromosomes, while the 2A and 2B chromosomes were fused in humans. Henceforth, the results were presented based on human chromosome set, given that we used the human reference for read mapping and variant calling above. Over the 23 chromosomal regions (22 autosomes and the X chromosome), there were 2,844 noncoding blocks (2,685 autosomal, 159 X-linked), 1,384 CNE blocks (1,324 autosomal, 60 X-linked), and 615 coding blocks (591 autosomal, 24 X-linked). The Y chromosome was analysed as one locus.

The species trees estimated in this block analysis on each chromosome are shown in figure 4.2 for the three sets of data. The 15 inferred trees were ordered according to their average posterior probabilities across all blocks (fig. 4.2). Trees 1 to 4 in figure 4.2 received genome-wide support $> 5\%$, identified as the maximum a posteriori (MAP) tree in 44%, 38.7%, 9.6% and 6.4% of the blocks, respectively. Together, these trees accounted for approximately 99% of all inferred trees across the genome.

Among the four trees, the unbalanced tree $((((S,T),E),V),P)$ (Tree 1 of fig. 4.2) and the balanced tree $(((S,T),(E,V)),P)$ (Tree 2 of fig. 4.2) had relatively high probabilities on each chromosome, with each supported by $\sim 40\%$ of the blocks across three sets of data. Tree 1 was indeed inferred in more blocks than Tree 2.

In this analysis, we formed blocks of 100 loci and inferred species tree from each block. The block

size should be sufficient to filter out the fluctuations in the coalescent process, so the differences among the blocks are primarily caused by heterogeneity in the migration rate along the genome. Among the three datasets, most of the noncoding blocks on autosomes had species tree distributions dominated by either Tree 1 or Tree 2. In contrast, there were signals of more diverse trees in coding and CNE blocks. It may be due to that the migration rate varies drastically in different types of data.

It is expected that one of the inferred trees should be the population phylogeny that reflects the history of population divergence, while the others are a result of gene flow. To identify the true population phylogeny, we obtained whole mitochondrial genome sequences (about 16,500 bps) for 43 bonobos (*P.p*), 18 N-C (*P.t.e*), 37 Western (*P.t.v*), 40 Eastern (*P.t.s*) and 59 Central chimpanzees (*P.t.t*), compiled by Lobon *et al.* (2016). The chimpanzee and bonobo reference sequences were removed as they may be chimeric. The reconstructed ML tree in figure S4.2a was clear of gene flow signatures and agreed with Tree 2 $(((S,T),(E,V)),P)$ of figure 4.2. The tree was also supported in the BPP analysis with a high posterior probability of 97.8%. ML analysis of the D-loop region (about 1100 bps) produced a very similar tree.

Based on the results, Tree 2 $(((S,T),(E,V)),P)$ was assumed to be the population tree in the subsequent analysis. It suggests a relationship of chimpanzee subspecies that the Nigeria-Cameroon (E) and Western (V) chimpanzee form a clade sister to that of the eastern (S) and central chimpanzee (T). The divergence time (τ) estimates under the MSC model in figure S4.1 suggest the split of E and V to predate that between the other pair of subspecies. The older divergence between E and V was consistent with previous evidence (de Manuel *et al.*, 2016; Hey *et al.*, 2018; Prado-Martinez *et al.*, 2013).

On the other hand, the trees of figure 4.2 hinted potential ancient migration between chimpanzee subspecies. For example, the high probability of 44% for tree $((((S,T),E),V),P)$ (fig. 4.2) might be indicative of certain migration between subspecies E and the ancestor ST. However, it is difficult to use the source of evidence to identify weak signals of gene flow, and there is no information on migration between sister species, which does not necessarily cause topological shifts. We then took a more methodical approach to building a model of gene flow.

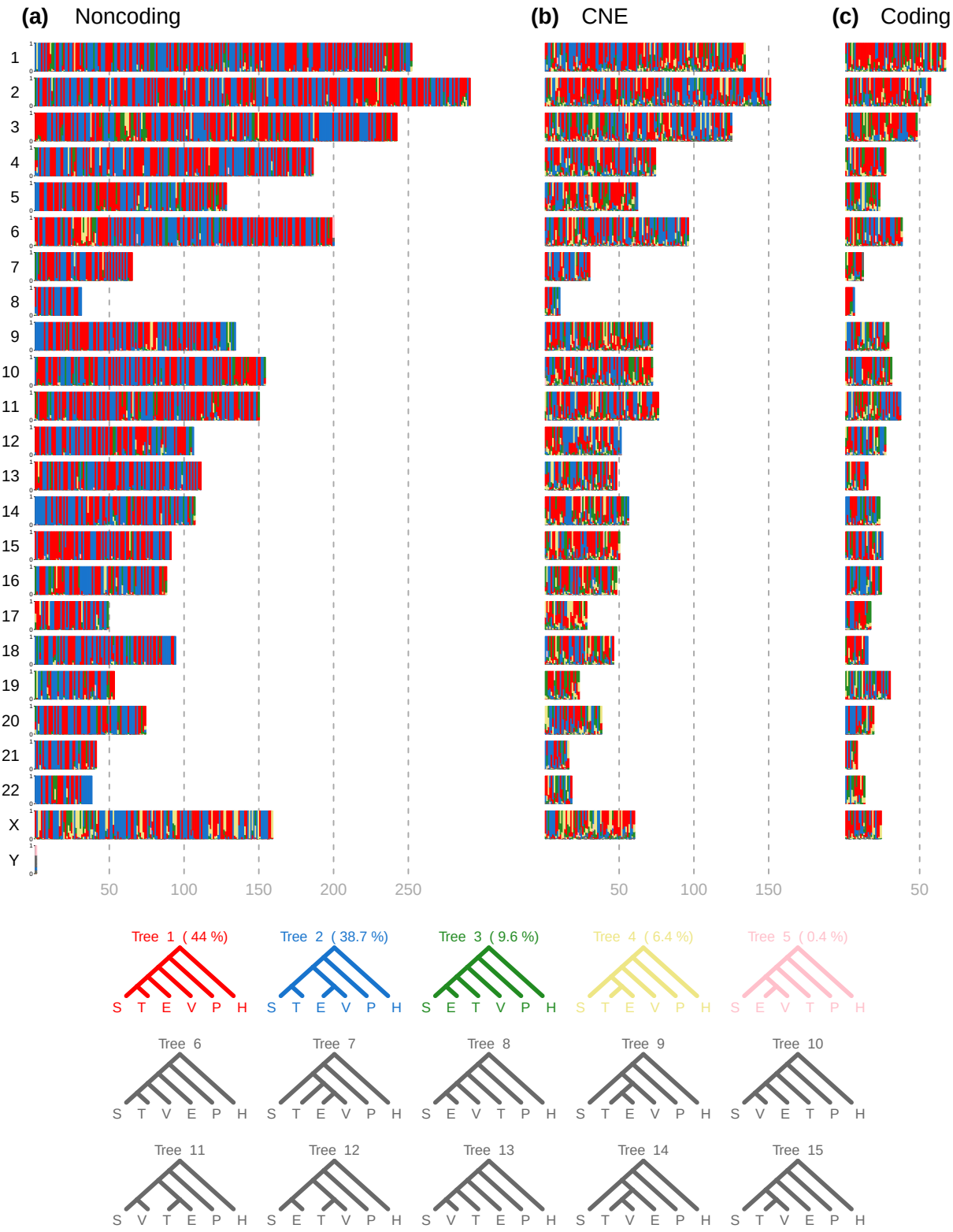

Figure 4.2: Posterior probabilities for species trees for the bonobo and the four chimpanzee subspecies in BPP analysis of 100-loci blocks under the MSC model with no gene flow. The height of each coloured bar represents posterior probability and ranges between 0 to 1. The five most probable trees are shown in colours, with the proportion of blocks across the three sets of data in which each tree was the MAP tree shown in parentheses. Tree 2 (38.7%) is consistent with the mitochondrial tree in figure S4.2a, used as the population phylogeny.

4.2.2 Construction of a model of gene flow for the chimpanzees and bonobos

In the first step of the model construction, we fitted migration models of three species using 3S and BPP. Bidirectional migration were specified between chimpanzee subspecies and between the ancestor and the bonobo. Each migration model includes 4 migration rates.

The migration rates and the results of hypothesis tests in the triplet analysis were summarized in figure S4.3a for 3S and S4.3b for BPP. In 3S, an upper bound of 1.5 is set for the migration rate ($M = Nm < 1.5$ migrants per generation). Nearly half of the rates estimated using 3S reached the upper bound. Both 3S and BPP suggested widespread gene flow between chimpanzee subspecies and between the two species, with an absence or low rate of migration into the western chimpanzee (V). Among the four chimpanzee subspecies, migration between the eastern (S) and central (T) chimpanzees is the most significant, as indicated by the highest migration rate M in BPP. The rates of other migration ranged approximately between 0.1 and 0.4, except those involving the western (V) as recipient (fig S4.3**b**). Additionally, migration from the chimpanzee to the bonobo was generally stronger than in the opposite direction.

In the next step, we formulated a joint migration model for each of the three data types, by including migration events supported by both 3S and BPP in the triplet analyses (table 4.1). The joint models for the CNE and coding data are the same, including 8 migration rates: STEV $\rightarrow$ P, P $\rightarrow$ STEV, ST $\rightarrow$ E, E $\rightarrow$ ST, V $\rightarrow$ ST, S $\rightarrow$ T, T $\rightarrow$ S, and V $\rightarrow$ E. The model for the noncoding data included one additional rate: ST $\rightarrow$ V. We then used BPP to fit the joint model for each data type to the 23 datasets for the 23 chromosomes, each of 500 loci. The rates averaged across chromosomes are shown in table 4.1. Migration with the strongest signal throughout the genome was between ST and E, with a rate of ~ 0.5 for ST to E, and an even higher rate in the opposite direction. Interestingly, gene flow between S and T, which was inferred to be most significant in the triplets, was found to be negligible under the joint model. Migration from V $\rightarrow$ E and from STEV $\rightarrow$ P was supported by coding data on several chromosomes and by noncoding data across nearly all chromosomes, but neither was detected on more than 5 chromosomes using CNE data. In the analysis under the joint models in table 4.1, we identified substantial evidence for four migration events (ST $\rightarrow$ E, E $\rightarrow$ ST, V $\rightarrow$ E and STEV $\rightarrow$ P) using noncoding and coding data, out of which migration from ST $\rightarrow$ E and from E $\rightarrow$ ST was also

Table 4.1: Average estimates of migration rates (M) and the number of significant tests in BPP analysis of datasets for the 23 chromosomes

Migration	Average rate	Significance
Noncoding, MSC-M model with 9 rates		
ST → E	0.54	23/23
E → ST	1.67	23/23
V → E	0.14	21/23
STEV → P	0.34	18/23
P → STEV	0.04	4/23
ST → V	0.05	3/23
V → ST	0.07	2/23
S → T	0.16	1/23
T → S	0.09	1/23
CNE, MSC-M model with 8 rates		
ST → E	0.52	19/23
E → ST	0.66	5/23
V → E	0.10	2/23
STEV → P	0.05	0/23
P → STEV	0.04	0/23
V → ST	0.19	3/23
S → T	0.09	0/23
T → S	0.09	0/23
Coding, MSC-M model with 8 rates		
ST → E	0.57	23/23
E → ST	0.75	13/23
V → E	0.09	2/23
STEV → P	0.14	6/23
P → STEV	0.04	1/23
V → ST	0.17	1/23
S → T	0.08	0/23
T → S	0.07	0/23

Note.— Twenty-three datasets (each of 500 randomly sampled loci) were constructed for the 23 chromosomes, and each dataset was analysed under the MSCI-M model with 9 or 8 migration rates. The column *significance* indicates the number of datasets in which the Bayesian test of gene flow is significant at the 1% level (i.e., $B_{10} \geq 100$).

well supported in CNE data. Kuhlwilm *et al.* (2019) identified an archaic ghost introgression from an extinct ape species into the bonobo. This was investigated by fitting the joint models including the ghost gene flow event (fig. S4.4) to the same data for the 23 chromosomes using BPP, and the results are shown in table S4.2. There was no evidence found for the ghost gene flow into bonobos.

We noted migration signals detected in the joint models were largely consistent with the results of Hey *et al.* (2018), based on 100 noncoding autosomal loci. Yet the model of Hey *et al.* (2018) includes migration between modern populations from E → S (with the rate $Nm = 0.011$), which was not detected in our analyses. In our parsimony approach, gene flow from E → S and E → T

was consolidated into a single migration event from E → ST. Thus we used the inferred model of migration in Hey *et al.* (2018), shown in figure 4.1b, in the subsequent genome-wide analysis. It includes five migration events, ST → E, E → ST, V → E, STEV → P and E → S.

Pilot runs were conducted to evaluate the model of Hey *et al.* (2018) together with three other models (fig. 4.3). Models *i* to *iv* were fitted using BPP to a randomly selected block of 200-loci per chromosome for each data type. This was to verify the model of gene flow after the E → S migration was added. The results largely confirm our results obtained earlier under the joint models (table 4.1). The E → S migration was significant, even in the presence of the E → ST gene flow. Also, the estimation of rates between species appeared to be affected by the specification of gene flow between chimpanzee subspecies. Specifically, there was an increase in the migration rate from STEV → P when gene flow between subspecies was incorporated in the model (fig. 4.3).

Selective pressures acting on coding and CNE loci may diminish the detectable signatures of gene flow, potentially due to the deleterious effects of such gene flow on fitness. Consequently, the observed migration rates were relatively low in comparison to those inferred from noncoding loci.

4.2.3 Estimation of migration rates and species divergence times

We estimated migration rates under the MSC-M model using blocks of 200 loci. A total of 1,422, 691, and 308 blocks were formed for the noncoding, CNE, and coding datasets, respectively. For each block, the MSC-M model in figure 4.1b was fitted, with 5 migration rates.

The estimated migration rates and the results of the Bayesian tests of gene flow for blocks on each of the 23 chromosomes were shown in figure 4.4. For each data type, the average migration rates are fairly consistent between different chromosomes (fig. S4.5). Noncoding regions generally exhibit higher rates of migration than exons and CNEs, while there are some exceptions suggested by gene flow V → E and E → S (fig. S4.6). This matches the observation in pilot runs.

The posterior means of the migration rates inferred from noncoding data were averaged to be ∼ 0.5 for $M_{ST\to E}$, and ∼ 1.5 for $M_{E\to ST}$ on each chromosome (fig. S4.5), and the rates were lower in coding and CNE blocks. They are the strongest migration detected, supported by the most evidence across these blocks (fig. 4.4). Specifically, we identified gene flow $ST \to E$ in 73.7% of all blocks and

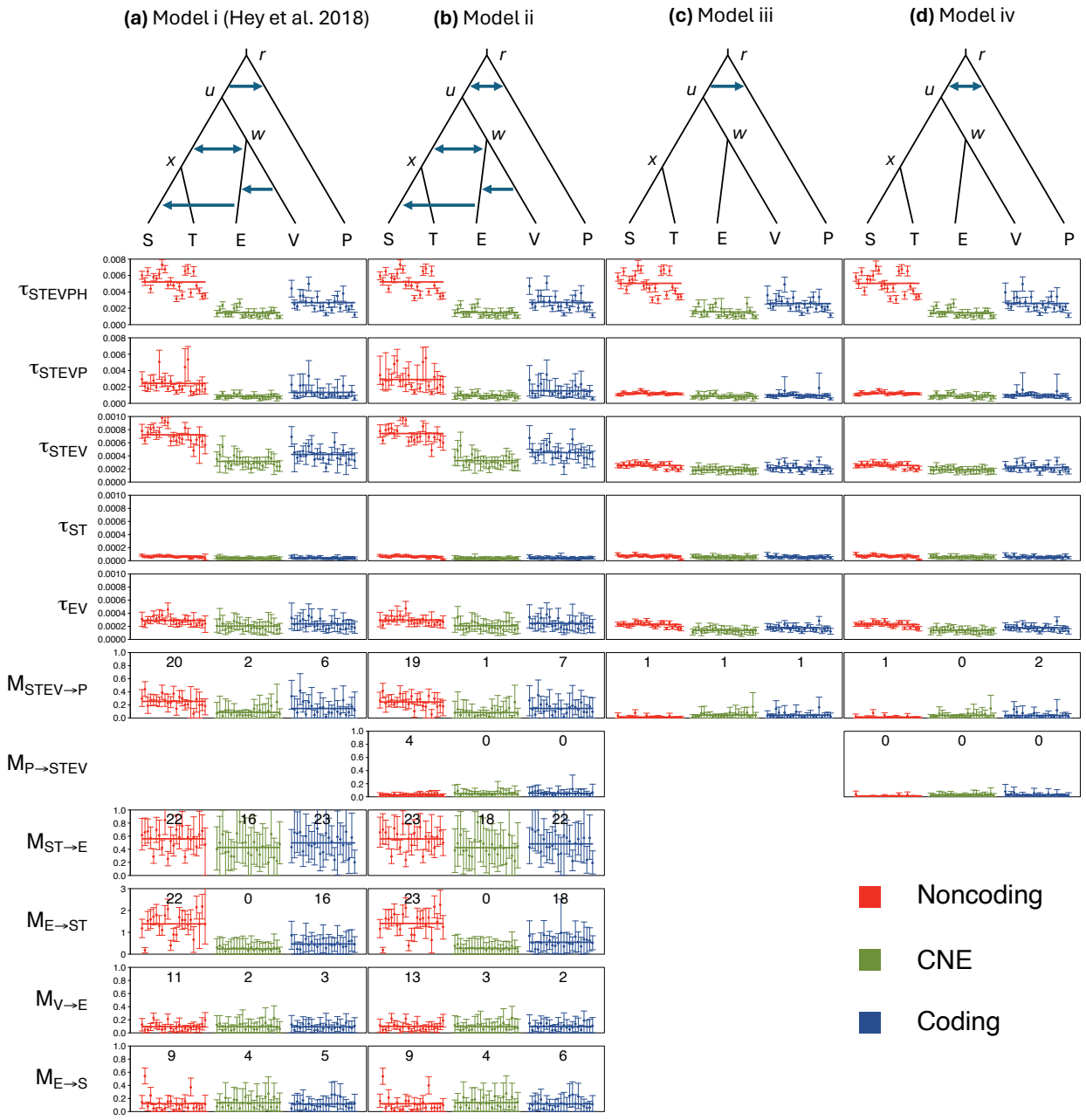

Figure 4.3: Migration (MSC-M) models tested in pilot runs using BPP. (**a**) Model i is the model inferred by Hey *et al.* (2018), with 5 migration rates. (**b**) Model ii extends Model i with an additional migration from P to STEV. (**c**) Model iii has a single migration from STEV to P. (**d**) Model iv includes bidirectional migration between the two species (between STEV and P). Outgroup H (humans) is included in the models but not shown here.

Below are posterior means and 95% HPD of population split times and migration rates ($M = Nm$) obtained in BPP analyses under the four MSC-M models above using 23 datasets of each type (noncoding, CNE or coding) for the 23 chromosomes, each dataset with 200 randomly selected loci. The numbers in the panels of migration rates (M) represent the number of datasets for each data type where the Bayesian test of gene flow is significant at the 1% level (i.e., $B_{10} \geq 100$)

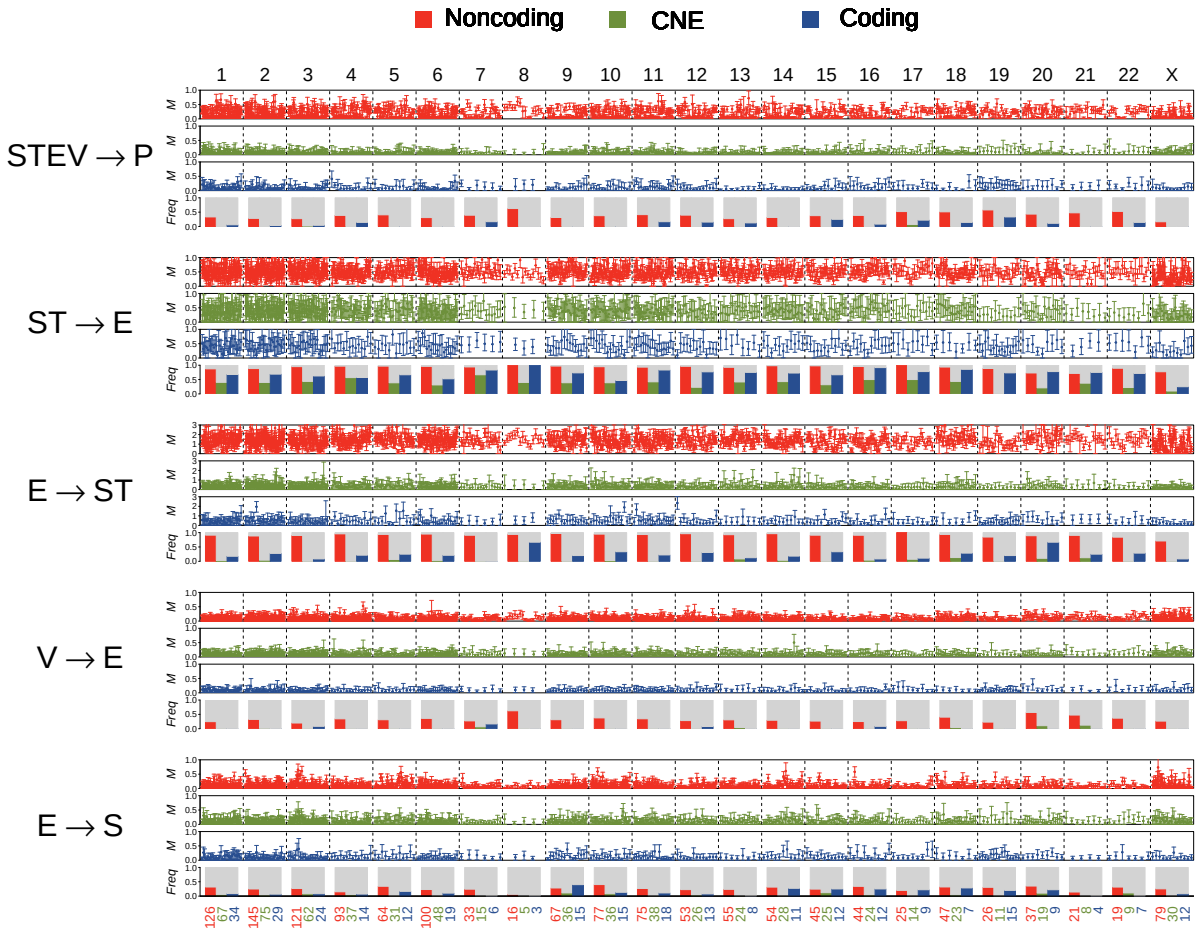

Figure 4.4: Posterior means and 95% HPD CIs for migration rates in BPP analysis of 200-loci blocks under the MSC-M model of figure 4.1b with 5 migration rates. The histograms display the distribution of Bayes factors calculated for blocks within each chromosome. The height of red/green/blue bars indicates the proportion of blocks in which rates are significant ($B_{10} > 100$) in noncoding/CNE/coding set, respectively. $B_{10} < 0.01$ are shown in black while $0.01 \leq B \leq 100$ are in grey. Bayes factors are calculated using the Savage-Dickey density ratio with $\varepsilon = 0.01$ (Ji *et al.*, 2023).

in more than 90% of noncoding blocks using the Bayesian test of gene flow. The significant rate for gene flow between ST and E explained the occurrence of tree $((((S,T),E),V),P)$ in the blockwise tree inference of figure 4.2, resulting in nearly half of the genome reflecting the history made up by the gene flow. The rate of the bidirectional migration appeared to be highly variable across blocks, and the rate variation along the genome was also predicted based on the colour patterns in figure 4.2.

Migration V → E and E → S between modern populations was estimated to be much weaker, with both rates around 0.1, and the difference was minor among the three data types and between different parts of genome. Despite the low rate from E to S, the reconciled tree $((((S,E),T),V),P)$ is

still represented in $\sim$ 10% of blocks.

In the BPP analysis under the MSC-M model, we detected considerable evidence for gene flow between ST and E. This event was also reported in recent studies incorporating all five populations (de Manuel *et al.*, 2016; Hey *et al.*, 2018). Further back, before the sequencing of Nigeria-Cameroon subspecies (E) by Prado-Martinez *et al.* (2013), gene flow had been identified between V and ST (Becquet *et al.*, 2007; Hey, 2010a; Wegmann and Excoffier, 2010). The evidence is indeed compatible with the migration model generated in our analysis (fig. 4.1b), where V $\rightarrow$ ST migration may occur through two gene flow events, V $\rightarrow$ E and E $\rightarrow$ ST, related by population E.

In addition to the migration among subspecies, gene flow between two species was also retained in the joint model and confirmed in the pilot runs above, with an assumed direction from the chimpanzee ancestor to the bonobo (STEV $\rightarrow$ P). The interspecific migration was estimated to have a rate of $\sim$ 0.15, averaged over blocks in the genome, significant in 501/1422 (35.2%) noncoding blocks, 27/308 (8.8%) coding blocks but almost none of CNE blocks. The estimation of the deep migration may suffer from low information content, even with 10 bonobo individuals included in the datasets. As indicated by the coalescent simulation (fig. S4.7), most bonobo lineages have coalesced before entering the chimpanzee ancestor at time τ_{STEV}, leaving only one to three sequences remaining at that time. In BPP, the estimation of migration rate for a given migration event under the MSC-M model relies on information such as the frequencies of lineages migrated through that event in genealogy samples, which is informative about the Poisson rate of migration events, $\frac{4M}{\theta}$. Estimates may be inconsistent due to large random error when there are too few lineages in the recipient population before the start of migration period. Our rate estimates appeared to be consistent and free of the issue.

The speciation times ($\tau = T\mu$) (fig. 4.5) and population sizes ($\theta = 4N\mu$) estimated under the MSC-M model of figure 4.1b on each chromosome were shown in figure S4.8. The ratio of mutation rates ($\frac{\mu_y}{\mu_x}$) between two types of data on the x- and y-axis of figure S4.8 is represented by the slope of linear regression in each panel. Exonic regions are expected to be conserved, while CNEs prove even more resistant to mutations, as indicated in the comparisons of parameters. The posterior means obtained using coding loci were proportionally smaller than those from noncoding loci, with a slope of 0.35 for τ (fig. S4.8a) and 0.75 for θ (fig. S4.8b), and slightly larger than the estimates of CNEs, with

a slope of 1.4 for τ (fig. S4.8a) and 1.04 for θ (fig. S4.8b). The linear relationship is well reflected for parameters of speciation times (τ) and modern population sizes (θ_{tip}), with a fitted $R^2 > 0.96$. This is due to that ancestral population sizes are more difficult to be precisely estimated.

We further investigated the impact of among-subspecies migration on the estimation of gene flow between species. One hundred 200-loci blocks were simulated using estimates from noncoding blocks, each with 10 haploid sequences for each chimpanzee subspecies and bonobos, and 4 for humans. The simulated data were then analysed under Models *i* to *iv* of figure 4.3, with results shown in figure 4.6. In the simulation analysis, when the migration between chimpanzee subspecies are not specified, as in Models *iii* and *iv*, there is clear underestimation of the migration rate $M_{STEV\to P}$ and the age of the chimpanzee root τ_{STEV}, similar to the observation in the pilot runs using real data (fig. 4.3). One possible explanation is that the time period for the between-species gene flow ($\tau_{STEVP} - \tau_{STEV}$) is overestimated due to the underestimation in τ_{STEV} if gene flow between subspecies is ignored. Given the amount of transferred alleles is stable through the period, the migration rate $M_{STEV\to P}$ is thus inferred to be lower in Model *iii* and *iv* than *i* and *ii*.

In the simulation, we also examined the model misspecification when chimpanzee subspecies E was unsampled or unavailable. We removed the sequences from that subspecies at each locus in the simulated data and analysed the reduced datasets under the 4-population model (except E) in figure S4.10, which includes 3 migration events STV $\to$ P, V $\to$ ST and V $\to$ S. The events V $\to$ ST and V $\to$ S were assumed among subspecies as substitutes for the gene flow from V $\to$ E, E $\to$ S, and that between E and ST, given that subspecies E is not in the model. Parameter estimates were shown in figure S4.10. $M_{STV\to P}$ were close to the true value of $M_{STEV\to P}$ in Model *i*, almost unaffected by the misspecified model ignoring E. Migration V $\to$ ST was inferred to have a rate of ~ 0.3, higher than $M_{V\to E} = 0.1$ but much lower than $M_{E\to ST} = 1.5$. Estimated migration rates from V to S were considerably low and not detected in any replicates using the Bayesian test of gene flow. This is likely because that the migration V $\to$ E and E $\to$ S was both very weak. Note that the migration rates used in the simulation are genome-wide averages. It is still possible for migration V to S to be detected in some genomic regions where V $\to$ E and E $\to$ S are stronger than the background level.

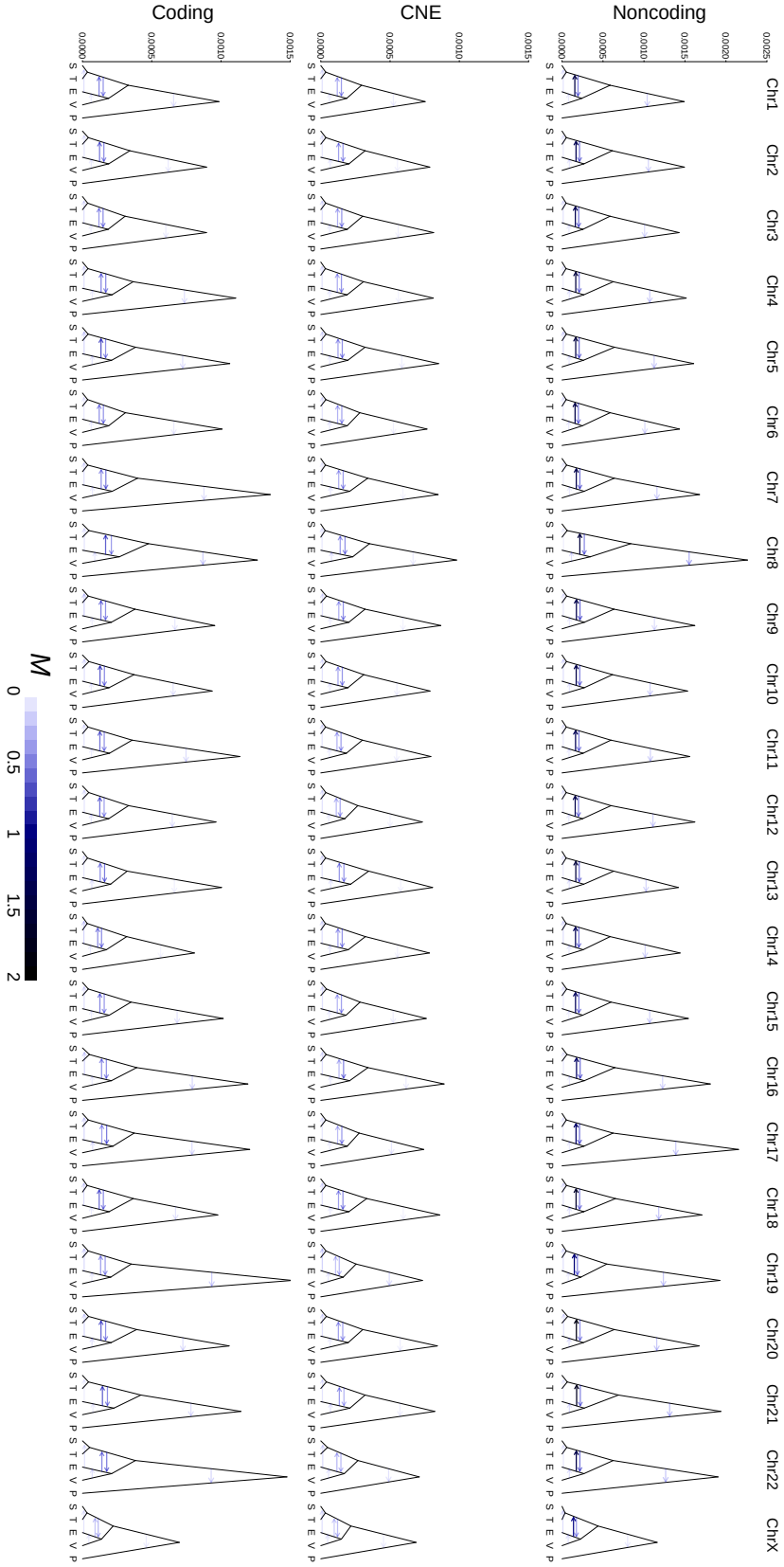

Figure 4.5: Migration history in each chromosomal region under the MSC-M model of figure 4.1. Speciation times (τ), represented on the y-axis, are calculated as the average of posterior means obtained using individual blocks in each chromosomal region. The intensity of the horizontal blue edges represents the five migration rates, which also have been averaged over blocks from each chromosome.

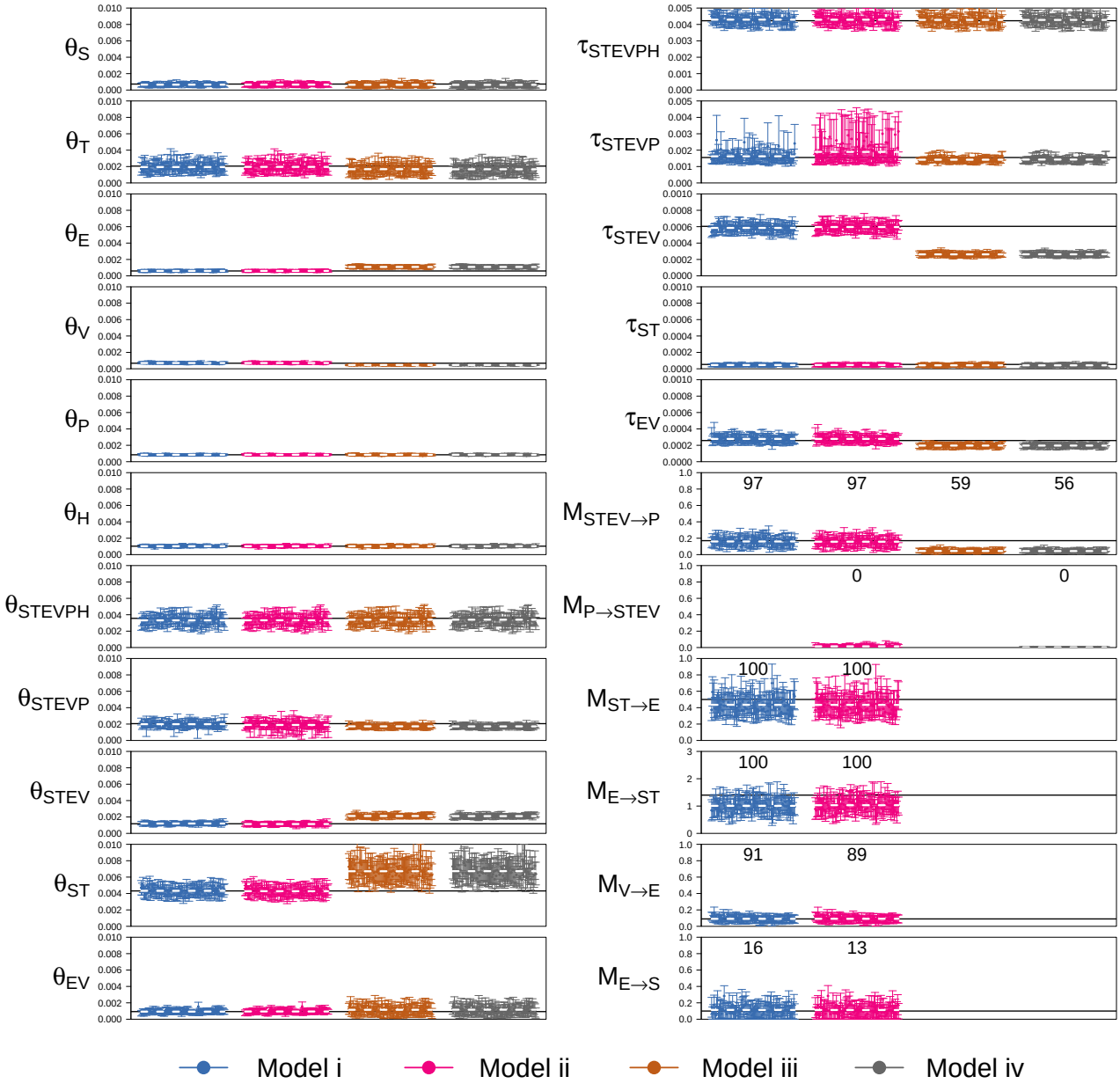

Figure 4.6: Posterior means and 95% HPD CIs for parameters in the MSC-M models of figure 4.3 in BPP analysis of 100 datasets simulated under Model *i*, each with 200 loci. The number above the CI bars is the number of replicates where the rate is significant in the Bayesian test of gene flow. True parameter values are shown using black solid lines. Averages of mean estimates across replicates are represented with white dashed lines.

4.2.4 Gene flow affecting the sex chromosomes

We then discuss the migration rates estimated on the two sex chromosomes, X and Y. We did not identify a different migration history on the sex chromosome X, while the rates $M_{STEV\to P}$, $M_{ST\to E}$ and $M_{E\to ST}$ were lower than those on autosomes (fig. S4.5). However, the male chromosome Y revealed a distinct pattern of gene flow.

In the species tree inference above, we identified a unique population phylogeny on the chromosome Y, $((((S,(E,V)),T),P),H)$ (fig. S4.2b and Tree 5 of fig. 4.2 &), consistent with the parsimony tree reconstructed in de Manuel *et al.* (2016). There were some examples of conflicting phylogenies for autosomes, Y chromosome and mitochondrial genomes (Chan *et al.*, 2012; Hallast *et al.*, 2016; Sarver *et al.*, 2021). The differences may be purely due to estimation artefacts. In our case, the ML tree in figure S4.2b has very short internals branches due to the accelerated coalescent process on the haploid chromosome, prone to systematic phylogenetic errors. There is considerable uncertainty in the tree inference using BPP on the chromosome. Hallast *et al.* (2016) attributed the tree to relatively slow coalescent in large populations T and the ancestor ST, while it was rarely supported in the analysis of autosomal and X chromosome data (fig. 4.2). It also cannot explain why there is no evidence of autosomes gene flow on Y. Providing that the patterns are real rather than a result of analytical artefacts, it may suggest a male-specific gene flow event between EV and S. Assume the population phylogeny $(((S,T),(E,V)),P)$, we fitted the MSC-M model with bidirectional migration between EV and S to the chromosome Y, which was treated as one locus in the analysis. The EV to S migration is significant and estimated to have rate $M_{EV\to S} = 0.46$, and migration rate in the opposite direction is close to zero.

In many primate species, including chimpanzees, male individuals are expected to be philopatric and stay in their natal groups, while females disperses between populations (Inoue *et al.*, 2008). Given that the chromosome Y is paternally inherited, the signature of gene flow via immigration of females may not be detected specifically on the Y chromosome. The evidence detected here may suggest the possibility of historic male-biased dispersal.

4.3 Discussion

Alongside our work, a wide variety of population genetics and phylogenomics methods have been applied to decipher the complex history of gene flow between chimpanzees and bonobos, revealing substantial signals between species and among subspecies, shown in table 4.2. We note that the differences in results may be explained by the data and methods used for studying gene flow. According to the comparative summary in table 4.2, the findings in previous studies are partially dependent on the populations considered in the models.

Table 4.2: Summary of methods, datasets and identified evidence in previous studies investigating gene flow in the *Pan* genus

		Detected gene-flow signal(s)	
Method or model	Data	Between species	Among subspecies
Brand *et al.* (2022)			
LEGOFIT	Genome-wide site patterns counts	P → ST	V → S
Kuhlwilm *et al.* (2019)			
FASTSIMCOAL[a]	Folded 3D-SFS (TVP)	Ghost → P, TV → P, P ↔ T	T ↔ V
ABC based on *S** statistics		Ghost → P	
Hey *et al.* (2018)			
IMA3	100 – 200 noncoding loci	STEV → P	ST ↔ E, V → E, E → S
de Manuel *et al.* (2016)			
D-STATISTIC[b]	SNPs across genome	P ↔ ST	ST ↔ E
FASTSIMCOAL[a]	Folded 4D-SFSs (STEP and STVP)	STE → P, STV → P, P ↔ ST	
TREEMIX	SNPs across genome	P → T	S → E
Prado-Martinez *et al.* (2013)			
D-STATISTIC[b]	Genome-wide site patterns counts		ST ↔ E, S ↔ E, S ↔ V
TREEMIX	SNPs across genome		E → S
Hey (2010a)[a]			
IM model of 2 populations	73 loci, mostly noncoding regions		T ↔ V
IM model of STV			V → ST, V → S
IM model of STVP		STV → P	V → ST, V → S, S ↔ T
Wegmann and Excoffier (2010)[a]			
ABC based on 96 statistics	265 microsatellites plus 26 intergenic loci	Fixed model assumed	
Becquet *et al.* (2007)[a]			
IM model of 2 populations	26 – 69 loci	P ↔ S	T ↔ V, S ↔ V, S ↔ T
Won and Hey (2005)[a]			
IM model of 2 populations	46 – 48 loci		V → T

[a]: Studies or analyses that did not incorporate all 5 populations.
[b]: Evidence from D-STATISTICS is interpreted as gene flow in both directions.

Early studies prior to Prado-Martinez *et al.* (2013) invoked a model-based approach but were limited to two populations and, most importantly, failed to accommodate the Nigeria-Cameroon chimpanzee (E), which was directly involved in 4 migration events in our model of figure 4.1b. As a recipient of gene flow, it received alleles from ST and V, while as a donor population, it sent alleles to ST and S. The transfer of genetic material along the path V → E → ST or V → E → S in the

MSC-M model is expected to be detected even in the absence of the intermediate population (fig. S4.10). Indeed there were detected signals supporting gene flow between V and S and V and ST in these studies (Becquet *et al.*, 2007; Hey, 2010a). Similarly, the gene flow V → S in Prado-Martinez *et al.* (2013) of table 4.2 was inferred based on the statistic $D(Human, V; S, T)$ without accounting for E. Given the current geographical distributions of the two subspecies (separate by nearly 3000 km), direct migration from Western (V) to Eastern chimpanzees (S) or the ancestor (ST) seems to be geographically implausible (Hey, 2010a). Instead, it is more likely that the migration was intermediated by another population in the middle, such as Nigeria-Cameroon chimpanzees.

Previous evidence of gene flow between S and T and V and T is likely an artefact caused by subsampling only two or three populations (Becquet *et al.*, 2007; Hey, 2010a; Won and Hey, 2005). We had similar noise in the analysis of triplet datasets of figure S4.3, which was rejected and excluded in the joint model (table 4.1). More recent analyses incorporating all subspecies have not found evidence for the events.

In this chapter, we identified robust evidence for multiple migration events using the full-likelihood method in BPP. Hey *et al.* (2018) applied a model-based approach in IMA3 and arrived at the same migration model as shown in figure 4.1b, but with lower rates for $M_{ST\to E}$ and $M_{E\to ST}$. The data employed in Hey *et al.* (2018) is equivalent to one or two blocks in the datasets compiled by us. Given the vast difference in data size, the rate differences are reasonable and a consensus is considered to have been reached. We then discuss the summary methods in table 4.2.

Methods such as TREEMIX and FASTSIMCOAL use allele frequencies summarized from SNPs across genome. The genome-wide averages are not expected to be informative for distinguishing a ghost introgression model from a model of gene flow between non-sister species (Pang and Zhang, 2024), and they may not be able to identify the direction of gene flow. For example, TREEMIX suggested migration from E to S in Prado-Martinez *et al.* (2013) and that in the opposite direction (S → E) in de Manuel *et al.* (2016), despite that the datasets used were largely overlapping. D-STATISTIC and LEGOFIT suffer from more serious information loss for pooling site pattern counts across the genome, which cannot be used to infer gene flow between sister species or directionality of gene flow between non-sister species. de Manuel *et al.* (2016) used D-STATISTICS and identified

gene flow between ST and E, indicated by evidence for gene flow between S and E and T and E, when examining different triplets. Migration between sister-species such as V to E and STEV to P were completely undetectable using these methods.

There are actually some limitations in our inference framework. First, the migration model of figure 4.1 was constructed using a two-step approach that formulates a joint model based on analysis of triplet data and revise the model to exclude false positive signals. The random subsets from each chromosome might not be sufficient to represent the entire chromosomal region. Signals not pronounced in the small subsets may be overlooked in our analysis.

Also, the parsimony assumption of gene flow events may not be correct. The migration between species was mostly inferred between P and ST and between P and STEV. Our analysis in triplets indicated stronger evidence of gene flow between P and ST than between P and EV (fig. S4.3), while they were assumed to occur between P and their ancestor STEV in the joint model. We have not yet statistically tested the possibility that they were two independent migration events in history.

Overall we suggest that insufficient sampling may be the major factor that prevent the early studies from correctly identifying the gene flow among subspecies. In particularly, the data for Nigeria-Cameroon chimpanzees (E) are critical since it was involved in all migration events among subspecies. Given the conflicting results produced in table 4.2, we highlight the consistency of full-likelihood methods in resolving gene flow history for complex scenarios.

4.4 Supplemental Information

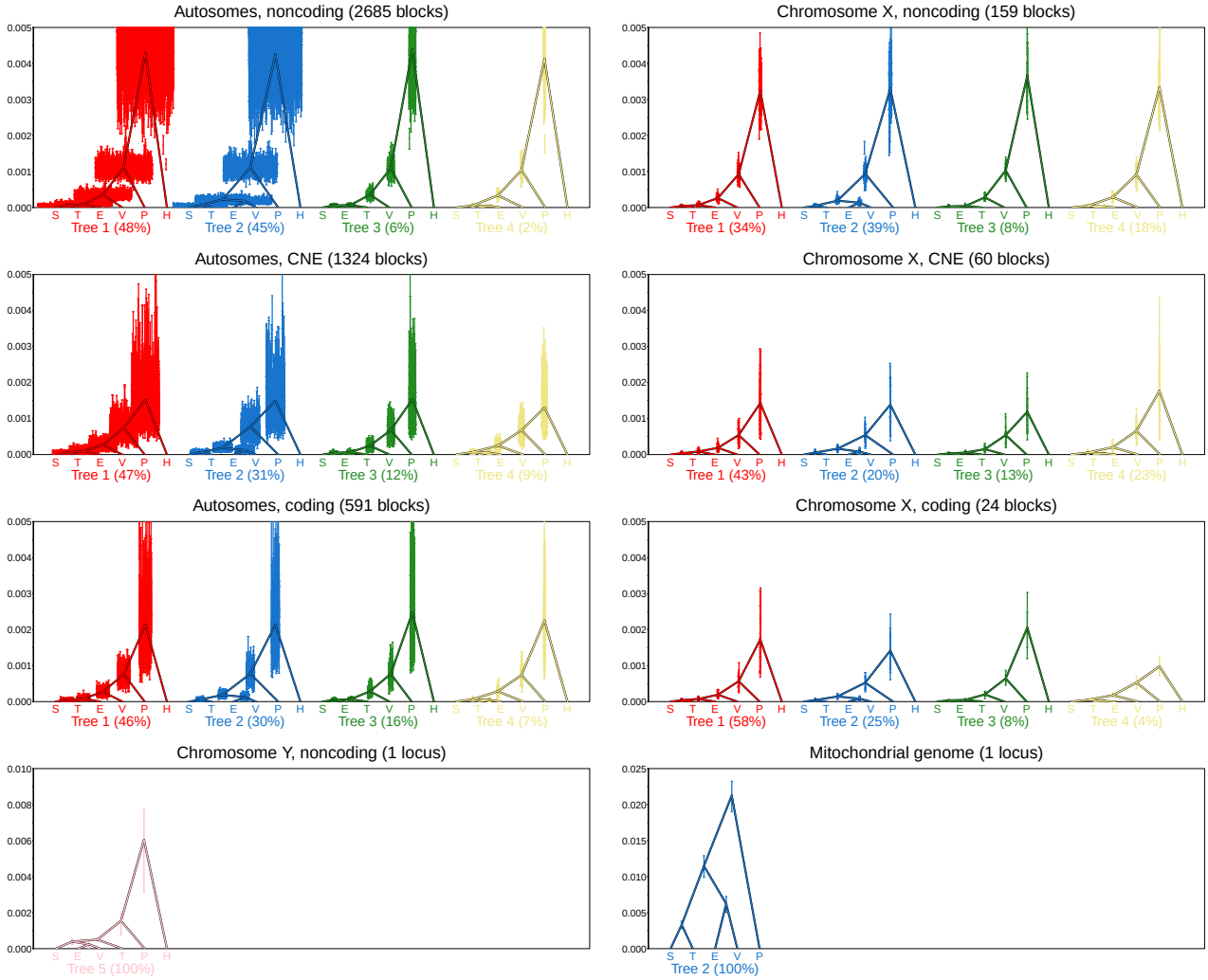

Figure S4.1: MAP species trees showing the 95% HPD CIs for species split times for noncoding, CNE, coding blocks on autosomes/X/Y/mitochondria, estimated under the MSC model with no gene flow (fig. 4.2). Blocks in which the MAP tree is not one of the top several trees are not shown here. The Y chromosome, 14 noncoding segments were concatenated and analysed as one locus. The whole mitochondrial genome was analysed as one locus. Node heights of the backbone trees reflect the averages across blocks. Species trees are coloured in the same way as in figure 4.2.

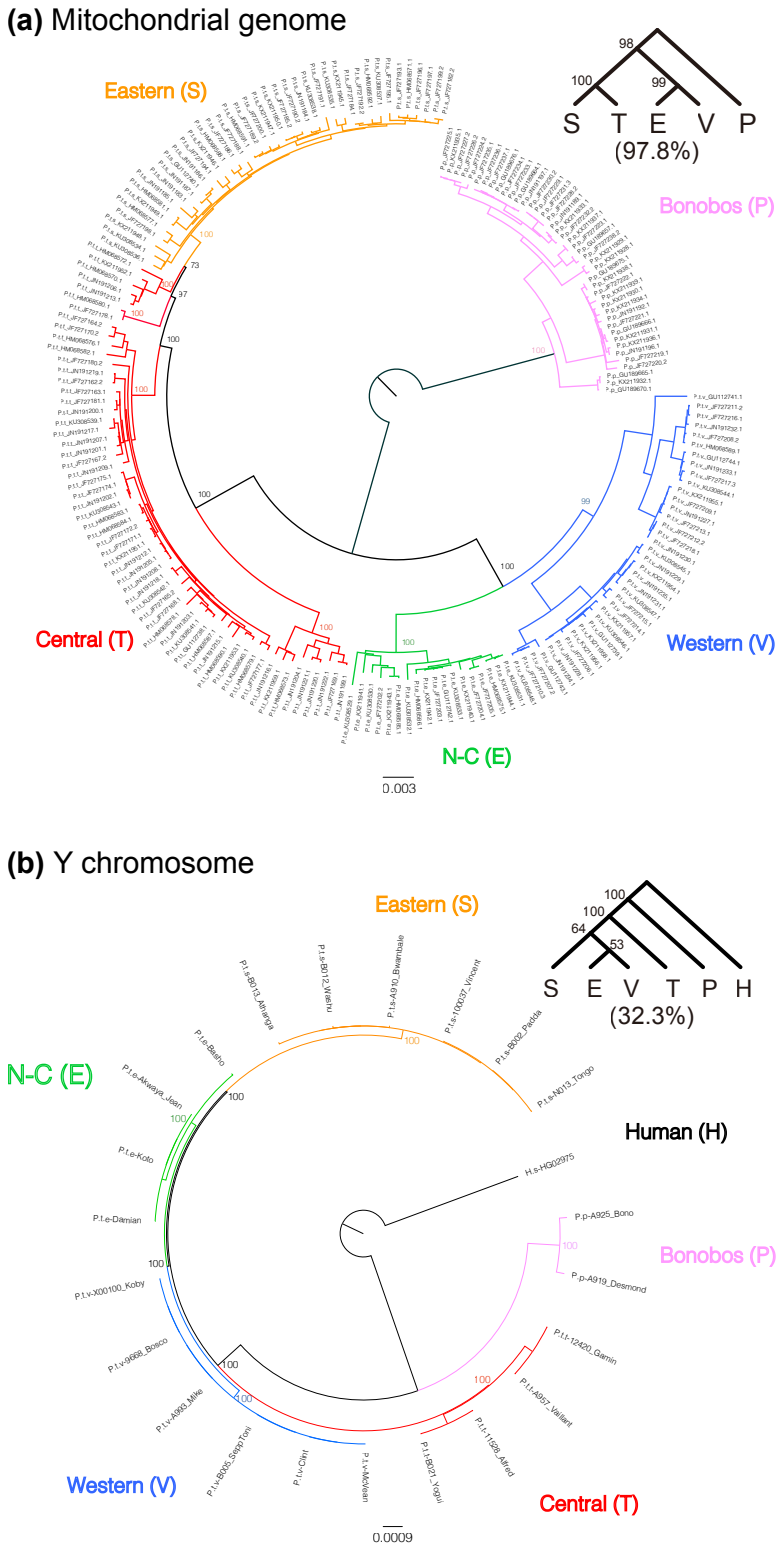

Figure S4.2: (**a**) ML tree inferred from RAxML analysis of the mitochondrial genome sequences. All subspecies except Central chimpanzees were monophyletic. (**b**) ML tree inferred from RAxML analysis of Y chromosome. The mitochondrial and Y species trees have posterior probabilities of 97.8% and 32.3% using BPP, with each treated as a single locus in the analysis.

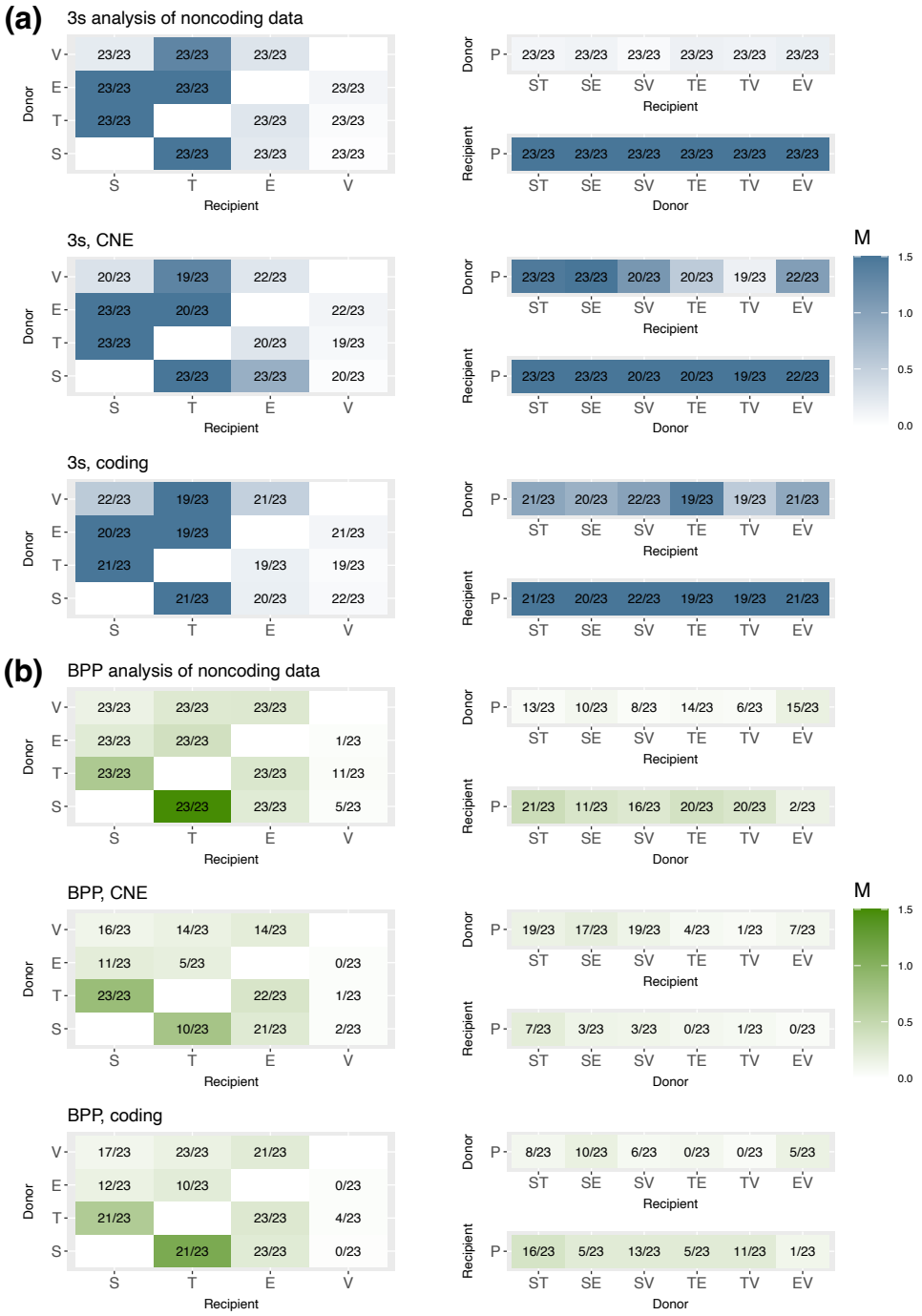

Figure S4.3: (**a**) ML (3S) and (**b**) Bayesian (BPP) estimates of migration rates ($M = Nm$) under the MSC-M model in analyses of triplet datasets for each chromosome. (**a**) In 3S analyses, for the CNE and coding data, the dataset for each chromosome includes all loci. For the noncoding data, 4000 loci were sampled at random if there were more than 4000 for the chromosome. Each dataset was analysed using 3S to fit the MSC-M model with four migration rates (that is, $M_{X\to Y}, M_{Y\to X}, M_{XY\to P}, M_{P\to XY}$ in the case of triplet XYP where X, Y are two chimpanzee subspecies and P is the bonobo). The intensity of color represents the migration rates averaged across datasets for each pair of populations, while the numbers in the cell (in the x/y format) record the number of datasets, out of 23, in which the rate passed the LRT at the 1% level. $M = 1.5$ is the upper limit set in the program. (**a**) In BPP analyses, for each of the three data types (noncoding, CNE, and coding), twenty-three datasets were constructed for the 23 chromosomes, each consisting of 500 randomly selected loci, and analysed using BPP under the MSC-M model with four migration rates.

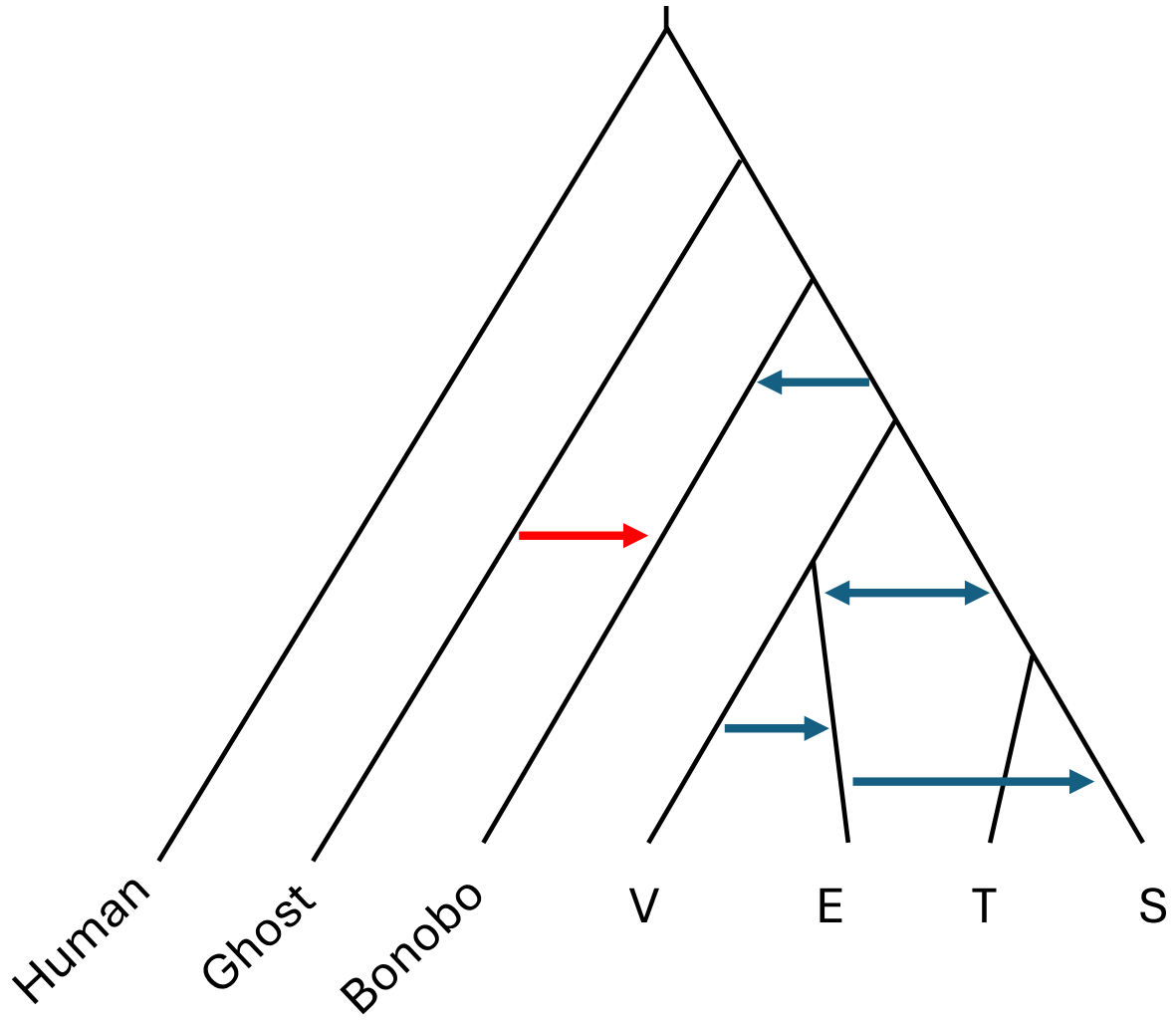

Figure S4.4: Migration model including five events in figure 4.1b and the ancient ghost event (Ghost → P) identified by Kuhlwilm *et al.* (2019), represented using a red arrow.

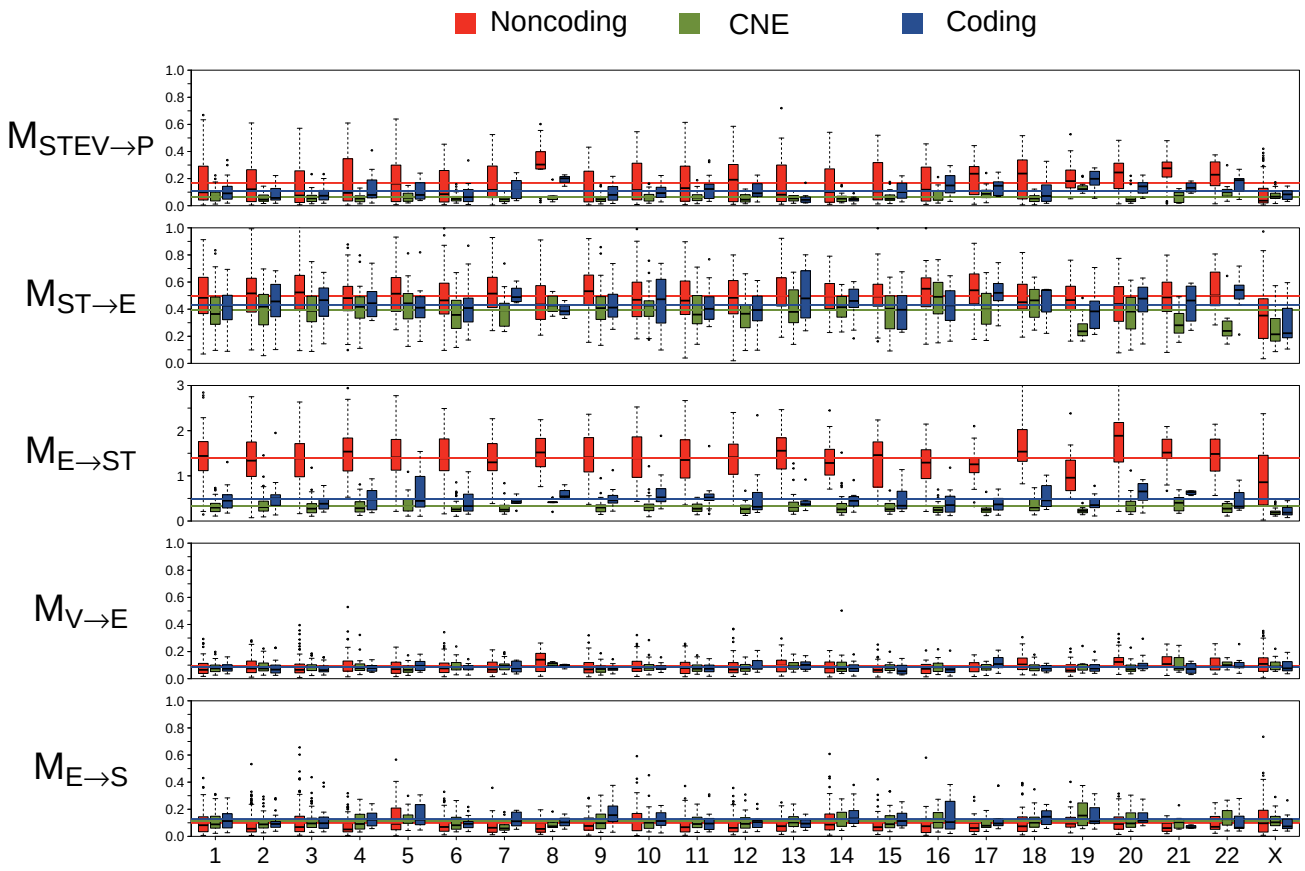

Figure S4.5: Boxplot of posterior means for migration rates in BPP analysis under the MSC-M model of figure 4.1b. The medians are represented by lines inside boxes, and the top and bottom edges of boxes indicate 25% and 75% quantiles, respectively. The whiskers represent the range of means for migration rates. The solid lines represent the genome-wide averages of migration rates (over all blocks).

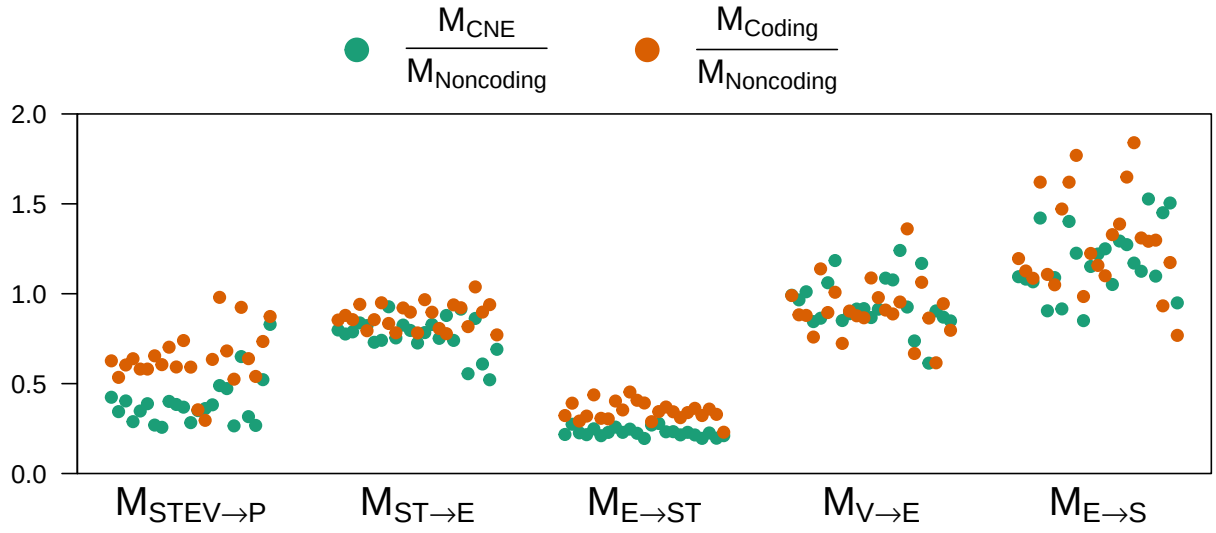

Figure S4.6: Ratios of average migration rates using the data of figure S4.5.

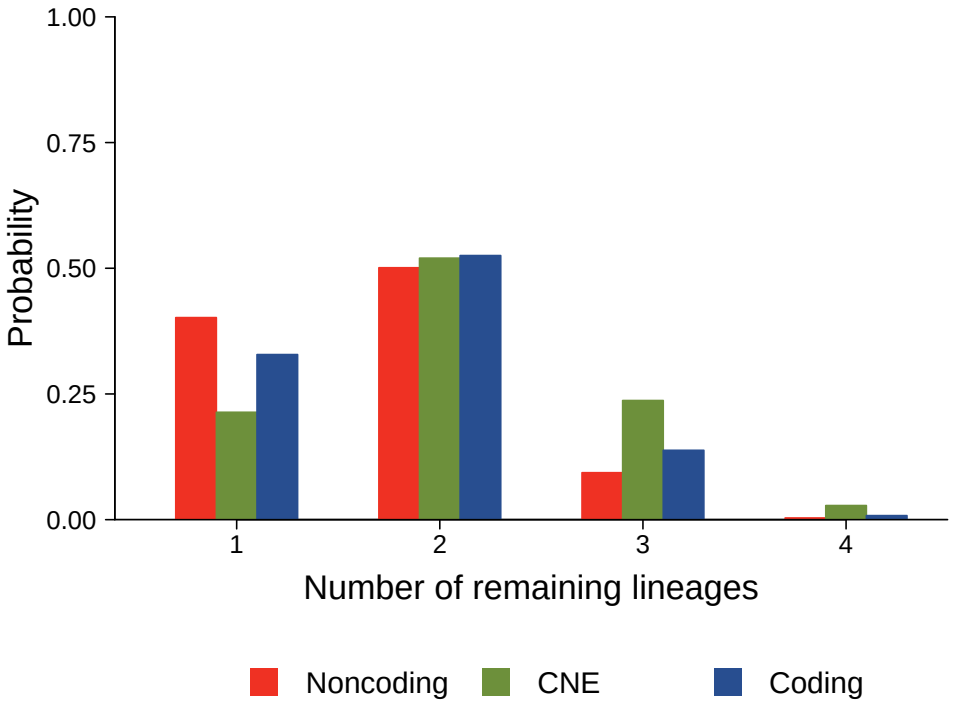

Figure S4.7: Number of bonobo sequences reaching the chimpanzee root (or τ_{STEV}), given 10 bonobo individuals (or 20 sequences) at each locus.

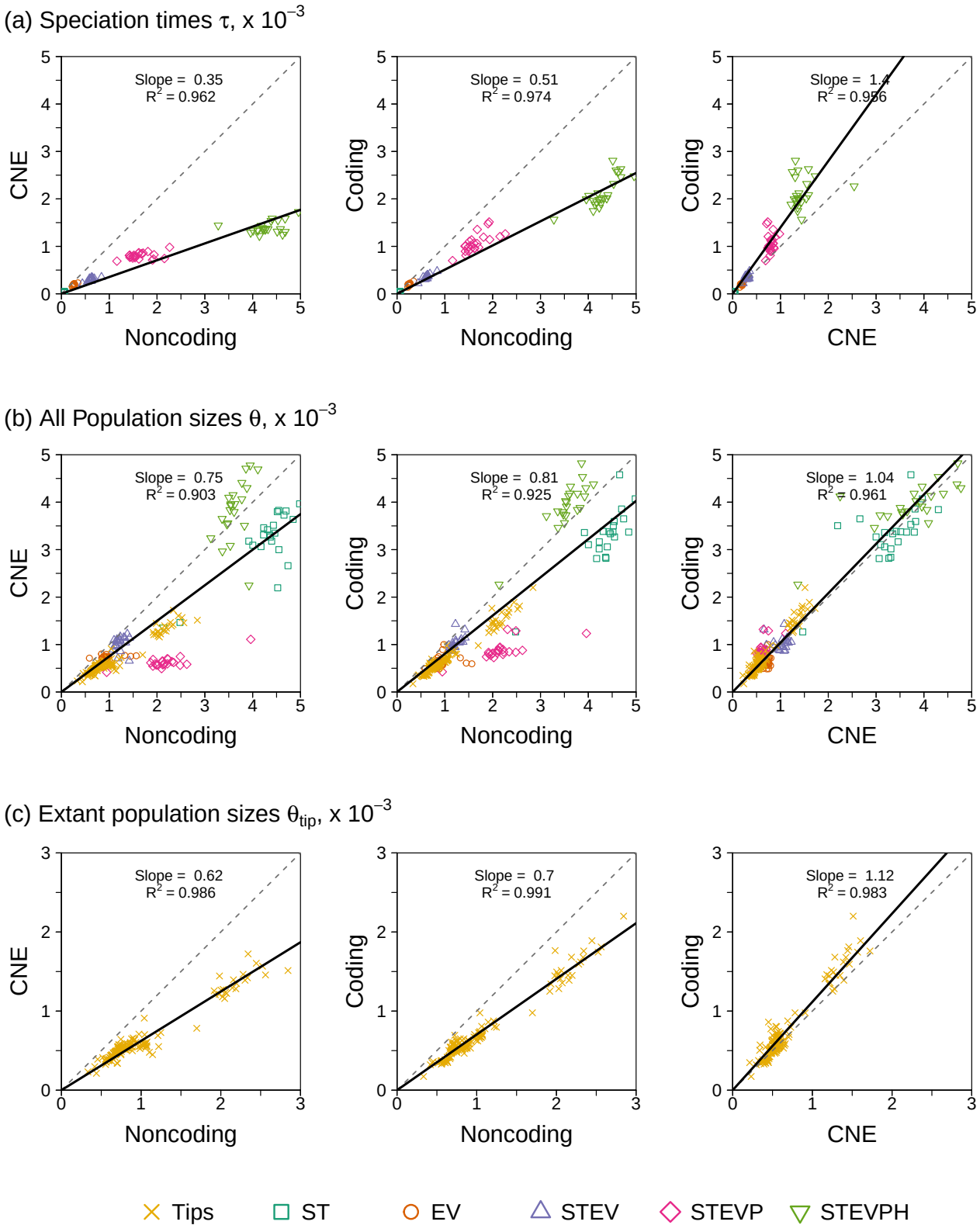

Figure S4.8: Estimates (posterior means) of τ and θ averaged over blocks on each chromosome for different data types under the MSC-M model of figure 4.1b. Estimates of τ and θ are multiplied by 10^3.

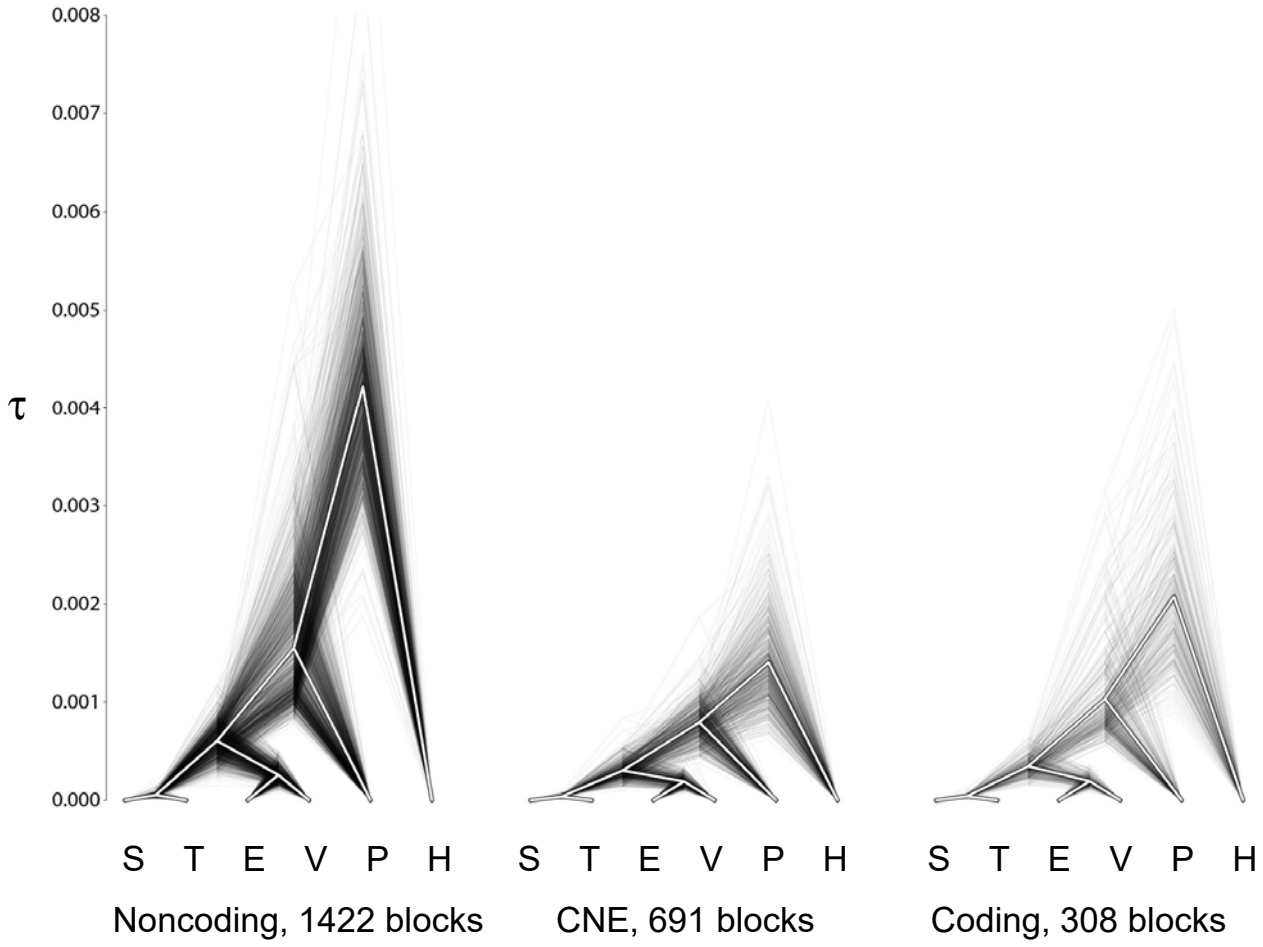

Figure S4.9: Densitrees from BPP analysis of the data blocks under the MSC-M model of figure 4.1b, showing the estimated species divergence times (τ). The backbone represents average speciation times over trees within each set.

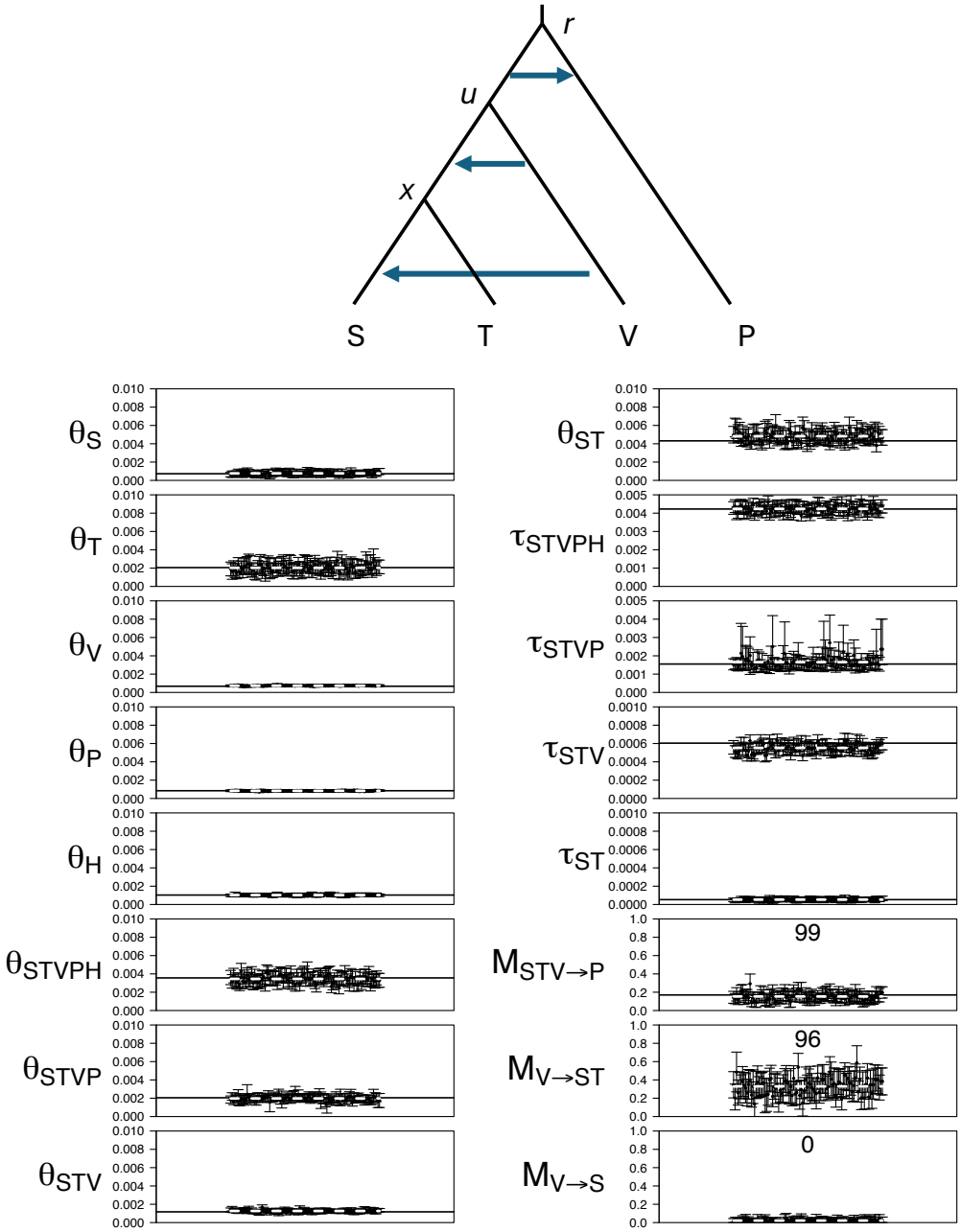

Figure S4.10: Migration model of 3 subspecies and bonobos, including 3 migration events: STV → P, V → ST and V → S. Posterior means and 95% HPD CIs for parameters in the MSC-M model in BPP analysis of 100 datasets simulated under Model *i* of figure 4.3 are shown below. This is the same data analysed in figure 4.6 but with sequences from subspecies E excluded at each locus. The number above the CI bars is the number of replicates where the rate is significant in the Bayesian test of gene flow. True parameters values of Model *i* are represented using black solid lines if these parameters also exist in the 4-population model. Averages of mean estimates across replicates are represented with white dashed lines

Table S4.1: The number of loci, the number of sequences per locus (median and range), the sequence length (median and range), and the number of variable sites for the three data sets

Dateset	# loci	# seqs	# sites	# variable sites
Coding	62,016	65 (5, 72)	170 (100, 2001)	3 (0, 103)
Noncoding	285,431	62 (4, 72)	876 (500, 2001)	24 (0, 202)
CNE	138,879	65 (6, 72)	166 (100, 2001)	3 (0, 68)

Note.– At most loci, there are more than 10 sequences, but the minimum is < 10 because we excluded male individuals for loci on chromosome X and subsampled some loci with many heterozygote sites.

Table S4.2: Average estimates of migration rates (M) and the number of significant tests in BPP analysis using models include a ghost gene flow event for the 23 chromosomes

Migration	Average rate	Significance
Noncoding, MSC-M model with 10 rates		
ST → E	0.59	23/23
E → ST	1.57	23/23
V → E	0.10	14/23
STEV → P	0.11	6/23
P → STEV	0.04	4/23
ST → V	0.06	3/23
V → ST	0.06	1/23
S → T	0.15	0/23
T → S	0.09	0/23
Ghost → P	0.00	0/23
CNE, MSC-M model with 9 rates		
ST → E	0.49	19/23
E → ST	0.68	6/23
V → E	0.11	4/23
STEV → P	0.04	0/23
P → STEV	0.04	0/23
V → ST	0.21	4/23
S → T	0.09	0/23
T → S	0.07	0/23
Ghost → P	0.00	0/23
Coding, MSC-M model with 9 rates		
ST → E	0.57	23/23
E → ST	0.73	12/23
V → E	0.08	1/23
STEV → P	0.10	1/23
P → STEV	0.04	2/23
V → ST	0.17	2/23
S → T	0.08	0/23
T → S	0.07	0/23
Ghost → P	0.00	0/23

Note.— For each data type, the fitted model consists of all migration events in table 4.1 and one ghost migration event from an unsampled ape lineage to the bonobo (Ghost → P). The column *significance* indicates the number of datasets in which the Bayesian test of gene flow is significant at the 1% level (i.e., $B_{10} \geq 100$).

Chapter 5

The Impact of Read Depth on Bayesian Analysis of Genomic Data under the Multispecies Coalescent Model

The multispecies coalescent (MSC) model accommodates the heterogeneity in genealogical relationships of sequences across the genome and provides a natural framework for analysis of phylogenomic data despite gene tree–species tree conflicts (Jiao *et al.*, 2021; Kubatko, 2019). Data suitable for analysis under the MSC are multilocus sequence alignments, or sequences that are short genomic fragments that are far apart. The large gap means that different loci have approximately independent coalescent histories, while intralocus recombination is unlikely in short genomic fragments. In such data, a locus is a short genomic segment and may not be protein-coding. Two strategies are commonly used to generate multilocus datasets in phylogenomic and population genomic analysis. The first is to sample short fragments that span hundreds to few thousand bps from sequenced genomes (e.g., Burgess and Yang, 2008; Dalquen *et al.*, 2017; Hey *et al.*, 2018; Thawornwattana *et al.*, 2018). For example, each segment may be 100-2000 bps long and separated by at least 2kb or 10kb. The second strategy is targeted sequence capture or reduced-representation sequencing, and includes RAD-seq (Eaton and Ree, 2013; Rubin *et al.*, 2012), ddRAD-seq (Ali *et al.*, 2016), exomes, transcriptomes, ultra-conserved elements (UCEs, Faircloth *et al.*, 2012), anchored hybrid enrichment (AHE, Lemmon *et al.*, 2012), conserved nonexonic elements (CNEEs, Edwards *et al.*, 2017), and rapidly evolving long exon capture (RELEC, Karin *et al.*, 2020), etc. This is a popular and less-costly alternative to whole-genome sequencing, widely used to generate phylogenomic datasets. The targeted genomic segments are typically 100-2000 bps long, and are sequenced to a high coverage.

Due to factors such as the cost, the coverage or read depth may not be very high, so that sequencing errors and genotype-calling errors may exist in the multilocus datasets, despite the common application of filters to remove or mask regions of low coverage to improve the quality of the sequence data (Thawornwattana *et al.*, 2018). In population genetics the impact of sequencing errors at different read depths has been studied extensively, with methods developed to correct biases in esti-

mates of population genetic parameters (such as the population size parameter $\theta = 4N\mu$) caused by sequencing errors at low read depths (Fumagalli, 2013). There does not appear to be any systematic study to examine the impact of sequencing errors at low coverage on phylogenomic inference under the multispecies coalescent (MSC) model, which addresses very different questions.

In this chapter we simulate multilocus genomic sequence data under the MSC model including sequencing errors at different read depths to examine the impact of genotyping errors on inference of species trees and estimation of population parameters in the MSC model with gene flow. We develop a Markov-chain model of read depths for sites along a sequence and simulate base-calling and genotype-calling errors in the sequence data. The data with genotype-calling errors are then analysed using the Bayesian program BPP to infer the species tree and to estimate parameters in the MSC-with-introgression (MSC-I) or migration (MSC-M), with the sequencing errors ignored, to assess the impact of sequencing errors on MSC-based inference.

5.1 Materials and Methods

5.1.1 Simulating sequence errors

We simulate multilocus alignments of unphased diploid sequences with sequencing errors by first generating correct sequences with no errors and then 'post-processing' the correct sequences to introduce genotype-calling errors using the base-calling error rate and simulated read depths for sites in the sequence in each sample. The procedures of the simulation are shown in the flowchart of figure 5.1. Note that here we do not simulate mapping errors or the use of filters to remove them. We do not simulate sites with zero coverage, as they are removed or masked during data processing if the neighbouring sites have high coverage.

We assume that all samples from the same species have the same average read depth and do not consider variable data qualities among samples of the same species. Because adjacent sites have a high chance of occurring in the same read, the read depths for different sites in the sequence at one locus are expected to be highly correlated. We develop a (hidden) Markov model to describe the transition of read depths at the adjacent sites in a sequence. Given the true genotype and the read

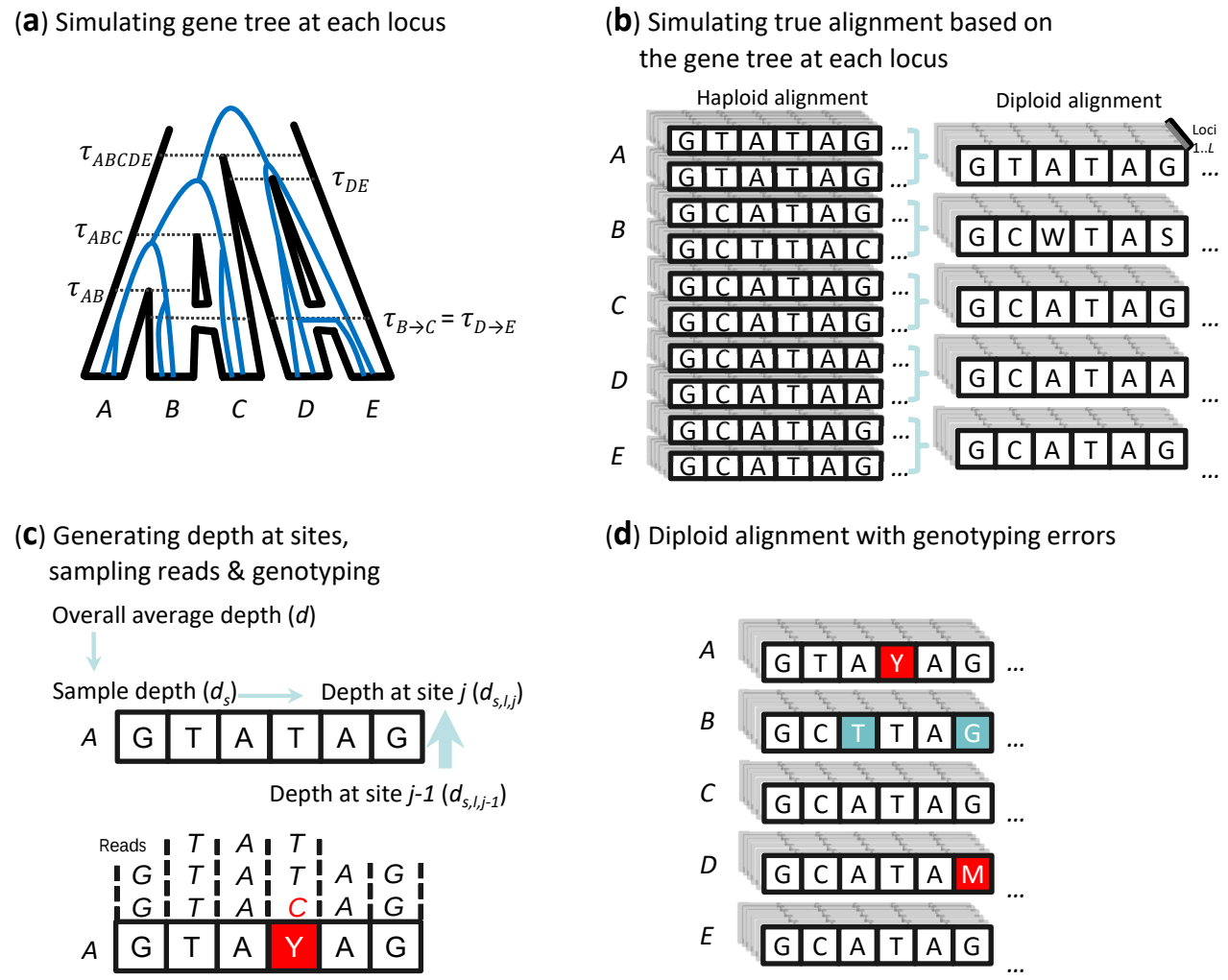

Figure 5.1: Simulation of multi-locus alignments of diploid sequences with genotyping errors at a given average read depth using BPP. (**a**) Simulation of gene trees at multiple loci under the MSC model with or without gene flow. (**b**) Simulation of true alignments using the gene trees (with two haploid sequences generated and merged into one diploid sequence). (**c**) Simulation of read depths using the beta model described in the paper, simulation of reads at sites in the sequence for each locus by binomial sampling of alleles and simulation of genotype calling by ML. (**d**) The resulting alignment of diploid sequences.
In (**c**), the diploid sequence from *A* is used as an example, with a base-calling error and a genotyping error shown in red. In (**d**), homozygotes miscalled as heterozygotes are in red while heterozygotes miscalled as homozygotes are in blue. Note that in the diploid sequence for *B* (**b**), the heterozygotes at two sites, W..S or (A/T)..(G/C), represent the haploid sequences A..G and T..C. Genotyping errors caused the heterozygotes to be mis-called as homozygotes TT and GG (**d**), and the resulting sequences with T..G at the two sites are chimeric and differ from the true sequences.

depth at each site, the reads at the site are generated through binomial sampling and are used to call the (observed) genotype by using maximum likelihood (Li, 2011).

5.1.2 A Markov model with a beta kernel for simulating read depths along a sequence

We use a pair of bounds for read depth: $d_{\min} = 2$, $d_{\max} = 100$, and use the beta distribution between those bounds to model the read depths for individual sites in a sequence. Let $x \sim \text{beta}(\alpha, \beta)$. This has mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$. Then

$$d = d_{\min} + x \cdot (d_{\max} - d_{\min}), \;\; d_{\min} < d < d_{\max}, \tag{5.1}$$

has the 4-parameter beta distribution, with parameters $(\alpha, \beta, d_{\min}, d_{\max})$. As the bounds $d_{\min}, d_{\max}$ are fixed,

$$x = \frac{d - d_{\min}}{d_{\max} - d_{\min}}, \;\; 0 < x < 1, \tag{5.2}$$

and d form a one-to-one mapping. We thus treat x as a scaled read depth and describe our model using x instead of d for simplicity.

Let $\bar{d}$ or $\bar{x}$ be the overall average read depth, specified in the simulation. Let $\bar{d}_s$ or $\bar{x}_s$ be the average read depth for species/sample s. We assume that all loci in all samples from the same species s have the same average read depth. This is generated as

$$\bar{x}_s \sim \text{beta}(\bar{x}a_s, (1-\bar{x})a_s), \tag{5.3}$$

where a_s is a parameter that describes how variable the average read depth is among species (with a larger a_s representing less variation).

Let d_{slj} be the read depth at the jth site in the lth locus in species/sample s. We use a (hidden) Markov model to simulate the transition of read depths at adjacent sites. For the first site ($j = 1$), we have

$$x_{sl1} | \bar{x}_s \sim \text{beta}(\bar{x}_s a_p, (1-\bar{x}_s)a_p), \tag{5.4}$$

with mean $\bar{x}_s$, where a_p describes how variable read depths are among positions (sites) of the same

sequence (with a larger a_p representing less variation). We used a_s = 500 and a_p = 1000, with more fluctuation between samples than between sites.

For site $j = 2, \cdots$, we generate x_{slj} from the beta distribution, $x_{slj} \sim \text{beta}(\bar{x}_{s,j}a_p, (1-\bar{x}_{s,j})a_p)$, but the mean $\bar{x}_{s,j}$ is given as a weighted average of the depth at the previous site and the mean depth for the species/sample:

$$\bar{x}_{s,j} = px_{s,l,j-1} + (1-p)\bar{x}_s \tag{5.5}$$

The parameter p controls how strongly correlated read depths are over adjacent sites. We used $p = 0.9$ based on an analysis of sitewise read depths in genomic sequence data (table 5.1, Model 1). The algorithm generates read depths $d_{sl1}, d_{sl2}, \cdots$ for sites in the sequence at locus l from sample s.

We considered an alternative model (table 5.1, Model 2), in which the read depth at the current site (x_{slj}) is assigned the read depth at the previous site ($x_{s,l,j-1}$) with probability p and generated from the beta distribution beta($\bar{x}_s a_p, (1-\bar{x}_s)a_p$) as in eq. 5.4 with probability $(1-p)$. This appeared to fit the empirical data less well (see below) and was thus not used.

Besides the beta kernel, we also considered an alternative Markov model for read depths based on the gamma kernel. The read depth at the current site j, d_{slj} is generated from the gamma distribution with the mean to be a mixture of the read depth at the previous site and the mean for the species/sample. The continuous gamma variables are rounded to integers and used as read depths. However, the algorithm may produce very low read depths (0 or 1), and truncation to apply the bounds $(d_{\min}, d_{\max})$ changes the mean read depth, making the model less attractive.

We processed real genomic data to collect the observed read depths at neighbouring sites to assess the fit of our models. Let f_{ij} be the observed frequencies of doublet sites with read depth i and j, respectively. Note that here we are assuming that the read depths along the sequence are Markovian, which we expect to be unrealistic but good enough for our purpose. The probability of observing two adjacent sites with read depths i and j under our model is

$$e_{ij} = p_i p_{ij}, \tag{5.6}$$

where p_i is the overall proportion of read depth i (or the stationary distribution of the Markov chain)

and p_{ij} is the transition probability (that is, the probability that the read depth for the next site is j given that the read depth for the current site is i). We can measure the discrepancy by

$$Q = \mathbb{E}(e_{ij} - f_{ij})^2 = \sum_i \sum_j e_{ij}(e_{ij} - f_{ij})^2. \tag{5.7}$$

5.1.3 Simulating genotypes given the read depth and true genotypes

Let ε be the base-call error rate. Given ε and the read depth d_{slj} we use the true genotype at the position to generate the reads by multinomial sampling. For each read, one of the two alleles at the position is chosen at random, and is then read correctly with probability $1-\varepsilon$ and incorrectly with probability ε. When a read error occurs, one of the three alternative bases is chosen at random. We will not deal with three or four alleles at one position and repeat the simulation for the site if more than two alleles occur.

The base-call error rate ε reflects the sequencing technology and may be independently estimated. Lou *et al.* (2013) mentions that Illumina sequencing machines of the time produce errors at a rate of 0.001–0.01, while the rate was estimated to slightly less than 10^{-3} for *Heliconius* genomes sequenced using paired-end reads on the Illumina Hi-Seq 2500 (Edelman *et al.*, 2019; Thawornwattana *et al.*, 2022).

Genotype calling. We assume that the base-call error rate ε is given, assumed to be the same among the reads, independent of the true base. Given the simulated reads, genotypes were called using maximum likelihood (ML) (Li, 2011). Given the data of k 1s and $(n-k)$ 0s among the n reads, where 0 refers to one allele and 1 the alternative allele, the likelihoods for the three genotypes (GT = 00, 01, and 11) are given by the binomial probabilities as

$$\begin{aligned} L(00|k) &= \mathbb{P}(k|\mathrm{GT}=00) = \tbinom{n}{k}(1-\varepsilon)^{n-k}\varepsilon^k, \\ L(01|k) &= \mathbb{P}(k|\mathrm{GT}=01) = \tbinom{n}{k}\left(\tfrac{1}{2}\right)^n, \\ L(11|k) &= \mathbb{P}(k|\mathrm{GT}=11) = \tbinom{n}{k}(1-\varepsilon)^k\varepsilon^{n-k}. \end{aligned} \tag{5.8}$$

The genotype achieving the highest likelihood is the called (inferred) genotype.

5.1.4 Implementation of the algorithm for simulating genotype-calling errors

The above algorithm for simulating site-wise read depths and for simulating diploid sequences with possible genotyping errors are implemented in BPP. The option variable `seqerr` in the control file has the following syntax:

```
seqerr = 5 0.01 500.0 1000.0 (read depth & base-calling error & a_samples & a_sites),
```

where the four parameters are the average read depth ($\bar{d}$), the base-calling error (ε), a_s, and a_p, respectively. In our simulation, we considered $\bar{d}$ = 3, 4, 5, 6, 8, 10, 15, 20, 25, 30, and $\varepsilon = 0$ (no error), 0.001, 0.005, 0.01.

The base-calling error rate ε is a fixed constant. To simulate a replicate dataset, we sample the average read depths ($\bar{d}$) for different species first. We then set up the alias method (Yang, 2014, p.421) for sampling reads given the read depth (which varies between $d_{\min}$ and $d_{\max}$) and ε. Then we loop through all loci, and introduce genotyping errors to the simulated diploid sequences with true genotypes, by sampling reads and calling genotypes by ML, before printing the alignments for each locus. In this algorithm, each average read depth ($\bar{d}$) is specific-specific, applied to all samples, all sequences, and all sites from that species in the whole dataset. However, $\bar{d}$ differs among species in the same dataset and among replicate datasets for the same species.

Heterozygotes in diploid sequences are coded using the International Union of Pure and Applied Chemistry (IUPAC) ambiguity codes (for example, Y stands for a T/C heterozygote). When the data are analysed by BPP, the `phase` control variable is used to instruct BPP to resolve each heterozygote genotype into the two alleles, averaging over all possible resolutions of phase at multiple heterozygous sites in the same sequence using the algorithm of (Gronau *et al.*, 2011; Huang *et al.*, 2022b).

5.1.5 Species tree estimation

This set of simulations examined the estimation of the species tree topology under the MSC model. We used the setting of Zhu *et al.* (2022). Multilocus sequence data were simulated assuming species trees B or U of figure 5.2. For tree B, the parameters were $\tau_r = 5\theta$, $\tau_s = 4.8\theta$, $\tau_t = 4.7\theta$, and

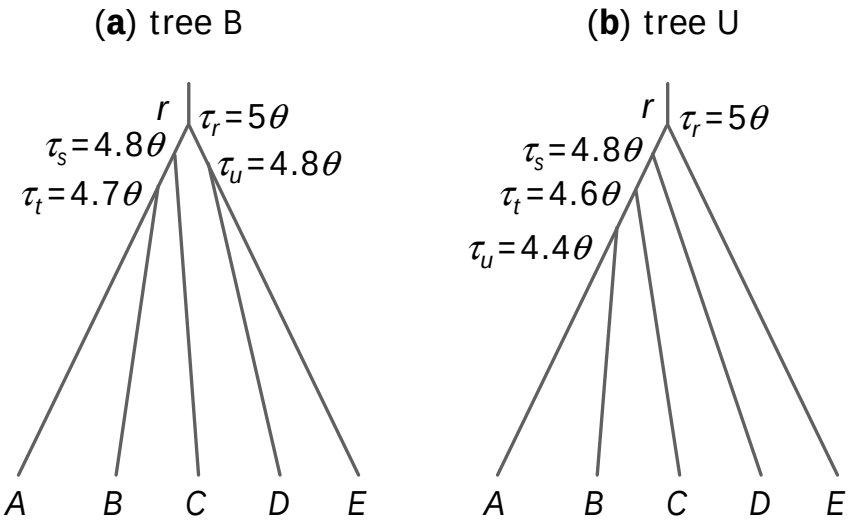

Figure 5.2: Species trees B and U for five species (A, B, C, D, E) used to simulate data for BPP estimation of the species tree (the A01 analysis). For balanced species tree B, the parameters are $\tau_r = 5\theta$, $\tau_s = 4.8\theta$, $\tau_t = 4.7\theta$, and $\tau_u = 4.8\theta$. For unbalanced species tree U, we used $\tau_r = 5\theta$, $\tau_s = 4.8\theta$, $\tau_t = 4.6\theta$, and $\tau_u = 4.4\theta$. In each tree, two values of θ are used: 0.0025 and 0.01. For analysis using ASTRAL and concatenation/ML, we included an outgroup species (O) with a divergence time of 10θ.

$\tau_u = 4.8\theta$. For tree U, they were $\tau_r = 5\theta$, $\tau_s = 4.8\theta$, $\tau_t = 4.6\theta$, and $\tau_u = 4.4\theta$. Two values were used for θ: 0.0025 and 0.01.

We generated either $S = 1$ or 4 diploid sequences per species per locus, with either $N = 250$ or 1000 sites in the sequence. Each replicate dataset consisted of $L = 40$ or 160 loci, with 5 or 20 unphased diploid sequences per locus. The number of replicate datasets was 100. The total number of simulated datasets, for all the combinations of tree, S, N, L, and θ is thus $2 \times 2 \times 2 \times 2 \times 2 \times 100 = 3200$.

Data were generated using the `simulate` option of BPP (Flouri *et al.*, 2018; Yang, 2015). Gene trees with branch lengths (coalescent times) were simulated under the MSC model (Rannala and Yang, 2003). Then sequences were "evolved" along the branches of the gene tree according to the JC model (Jukes and Cantor, 1969), and the sequences at the tips of the gene tree constituted data at the locus.

Each dataset was analysed using BPP to estimate the species tree. The subtree-pruning-and-regrafting (SPR) algorithm was used to move between species trees (Flouri *et al.*, 2018; Rannala and Yang, 2017). We integrated out θs analytically through the use of the conjugate inverse-gamma priors (Flouri *et al.*, 2018), which may help with MCMC mixing. We assigned inverse-gamma (IG)

priors to parameters in the MSC model: $\theta \sim \text{IG}(3, 2\theta)$ for population size parameters and $\tau_0 \sim \text{IG}(3, 10\theta)$ for the age of the root, with the mean matching the truth. As the starting species tree affects the time taken to reach stationarity, but not the mixing efficiency of the Markov chain after the burn-in, we used the true species tree as the starting tree. We calculated the posterior probabilities for the species tree and clades to measure performance.

We also analysed the data using ASTRAL and concatenation/ML to estimate the species tree. The sequence data included an outgroup species (O) to root the tree, which diverged from the ingroup species at time $\tau = 10\theta$ (fig. 5.2). For ASTRAL analysis, we used RAXML to reconstruct the gene tree for each locus under the JC model and then used ASTRAL to generate the species tree. The RAXML analysis treats the diploid sequence with heterozygotes as a haploid sequence with ambiguities; for example a T/C heterozygote is treated as either T or C (Andermann *et al.*, 2019; Huang *et al.*, 2022a). The concatenation method is applied to the case of $S = 1$ diploid sequence per species only. Sequences from all loci are concatenated and the super-alignment is analysed using RAXML to generate one tree, which is the estimate of the species tree. Again heterozygotes are treated as ambiguities.

5.1.6 Estimation of divergence times, population sizes, and rates of gene flow

In this set of simulations, we examined the estimation of parameters in the MSC model with gene flow, such as the species divergence times (τ), population sizes, and the rates of gene flow (φ or $M = Nm$), with the species tree fixed. We assumed species trees B or U of figure 5.3, and for each tree, we used two models of gene flow: MSC-I (Flouri *et al.*, 2020) and MSC-M (Flouri *et al.*, 2023), each with two unidirectional gene-flow events. We used two values for the population-size parameter θ: 0.0025 and 0.01. For tree B, the divergence times were $\tau_r = 5\theta$, $\tau_s = 4\theta$, $\tau_t = 3\theta$, and $\tau_u = 4.5\theta$. For tree U, the divergence times were $\tau_r = 5\theta$, $\tau_s = 4\theta$, $\tau_t = 3\theta$, and $\tau_u = 2.5\theta$. The introgression times under the MSC-I model were $\tau_b = \tau_c = \theta$, and $\tau_d = \tau_e = \theta$. In the MSC-I model, $\varphi_{bc} = 0.3$ and $\varphi_{de} = 0.2$, while in the MSC-M model, we used $M_{bc} = m_{bc}N_C = 0.3$ and $M_{de} = 0.2$.

For each parameter setting, we generated 100 replicate datasets. Each dataset consisted of $L = 40$ or 160 loci, with $S = 1$ or 4 diploid sequences per species at each locus, with the sequence length to be either $N = 250$ or 1000 sites. In total 3200 datasets were generated.

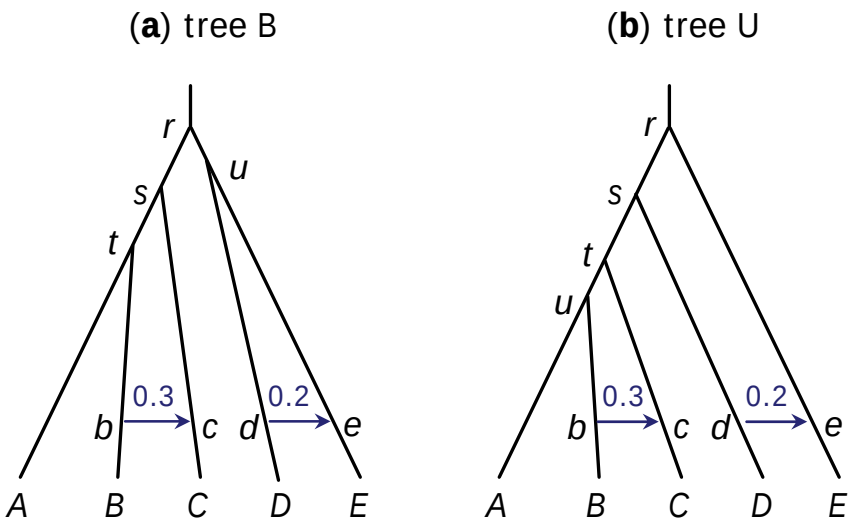

Figure 5.3: Two species-tree models of gene flow used in the simulation to evaluate Bayesian parameter estimation. Gene flow is modeled using either MSC-I or MSC-M. The parameters for tree B are $\tau_r = 5\theta$, $\tau_s = 4\theta$, $\tau_t = 3\theta$, $\tau_u = 4.5\theta$, $\tau_b = \tau_c = \theta$, and $\tau_d = \tau_e = \theta$, while those for tree U are $\tau_r = 5\theta$, $\tau_s = 4\theta$, $\tau_t = 3\theta$, $\tau_u = 2.5\theta$, $\tau_b = \tau_c = \theta$, and $\tau_d = \tau_e = \theta$. We used two values for θ: 0.0025 or 0.01. In the MSC-I model, we used $\varphi_{bc} = 0.3$ and $\varphi_{de} = 0.2$, while in the MSC-M model, we used $M_{bc} = m_{bc}N_C = 0.3$ and $M_{de} = 0.2$.

Each replicate dataset was analysed using BPP v.4.7 (Flouri *et al.*, 2018) to estimate the 21 parameters in the MSC-I model or 15 parameters in the MSC-M model. The correct species tree and the correct model (JC) were assumed. Under the MSC-I model, we assumed the same θ parameter for a branch on the species tree before and after an introgression event (by specifying `theta-model=linked-msci` in BPP control file), ensuring that $\theta_b = \theta_B$, $\theta_c = \theta_C$, $\theta_d = \theta_D$, $\theta_e = \theta_E$. Gamma priors were assigned on the population size parameters (θ) and the age of the root on the species tree ($\tau_0 = \tau_r$), with the shape parameter 2 and the prior means equal to the true values: $\tau_0 \sim \text{G}(2, 160)$ and $\theta \sim \text{G}(2, 800)$ for $\theta = 0.0025$, and $\tau_0 \sim \text{G}(2, 40)$ and $\theta \sim \text{G}(2, 200)$ for $\theta = 0.01$. The introgression probabilities under MSC-I were assigned the prior beta$(1, 1)$, while the migration rates under MSC-M are assumed the prior $M \sim \text{G}(1, 10)$. While the same θ was used for all species on the species tree in the simulation, every branch on the species tree had its own θ when the data were analysed using BPP.

We used 32,000 iterations for burnin, after which we took 10^5 samples, sampling every 2 iterations. Analysis of each dataset took ≈ 4 hours on a single thread for small datasets of 40 loci and 10 sequences per locus or ≈ 23 hours for large datasets of 160 loci and 40 sequences per locus.

5.2 Results

5.2.1 Empirical examination of read depths at adjacent sites in the genome

Table 5.1: Deviation measuring goodness of fit (Q, eq. 5.7) of two models for read depths along the sequence to observed data from two sequenced genomes (fig. 5.4)

	Chimpanzee genome (20.26X)		Rabbit genome (4.21X)	
p	Model 1	Model 2	Model 1	Model 2
0	0.0023	0.0023	0.0626	0.0626
0.1	0.0022	0.0025	0.0631	0.0695
0.2	0.0021	0.0033	0.0618	0.0779
0.3	0.0019	0.0049	0.0586	0.0879
0.4	0.0017	0.0072	0.0533	0.0995
0.5	0.0014	0.0106	0.0457	0.1128
0.6	0.0011	0.0150	0.0358	0.1277
0.7	0.0008	0.0206	0.0237	0.1444
0.8	0.0005	0.0275	0.0113	0.1628
0.9	0.0002	0.0360	0.0035	0.1830

Note.– Model 1 is the beta model based on eq. 5.5. Model 2 is the alternative model based on the beta distribution described in the text. Model 1 is used in our simulation as it fits the empirical data better.

We used site-wise read depths in sequenced genomes to assess the goodness of fit of our Markov-chain models of read depths along the sequence. The proportions (f_{ij}) of site doublets with read depths i, j for a high-coverage chimpanzee genome (average coverage 20.26X) sequenced by Prado-Martinez *et al.* (2013) and a low-coverage rabbit genome (4.21X) by Andrade *et al.* (2024) were used to generate empirical estimates of transition probabilities (fig. 5.4). These suggest strong correlation in read depth between adjacent sites, with very high probabilities that the read depth for the next site will be identical or very similar to that for the current site. We fitted the two beta models described above to the two depth datasets, and the goodness of fit for two models is measured using the sum of squared differences, Q (eq. 5.7). The results (table 5.1) suggest that the conditional-mean model fitted the data better for both datasets, with the parameter $p = 0.9$. This model is used in our study to simulate read depths along the sequence, given the average read depth.

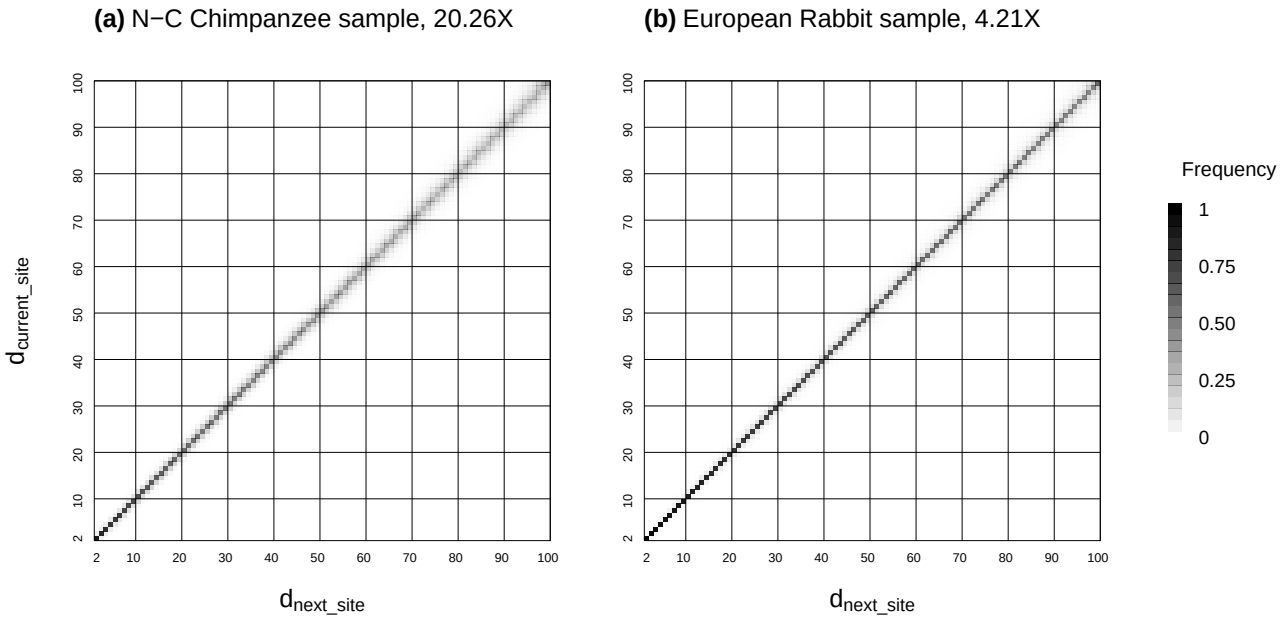

Figure 5.4: Heat-map representation of empirical transition probabilities of read depths at two adjacent sites ($\hat{p}_{ij} = f_{ij}/f_i$) estimated from (**a**) a Nigeria-Cameroon chimpanzee genome (*Pan troglodytes ellioti*), sequenced on the Illumina HiSeq 2000 platform to an average depth of 20.26x by Prado-Martinez *et al.* (2013) (NCBI accession: SRX360475) and (**b**) a European rabbit genome (*Oryctolagus cuniculus*), sequenced on the Illumina NovaSeq 6000 platform to an average depth of 4.21x by Andrade *et al.* (2024) (SRX21096756). The shading in each cell represents the frequency (f_{ij}) that the next site has read depth j given the read depth at the current site (i). Each row sums to 1.

5.2.2 Species tree estimation in presence of genotyping errors

The probabilities of inferring the correct tree using BPP, ASTRAL and concatenation/ML are summarized in figure 5.5. The average posterior probabilities for the true tree are shown in figure S5.3.

First, we consider the standard BPP analysis (BPP in fig. 5.5), treating the sequences as diploid sequences, averaging over all possible phase resolutions at heterozygote sites (Gronau *et al.*, 2011; Huang *et al.*, 2022b). Note that in our simulation the data size is fixed, and the results obtained when there was no sequencing errors ($\varepsilon = 0$) constitute the best-case scenario and provide a reference for comparison. Note that in the smallest datasets (with $L = 40$ loci, $S = 1$ diploid sequence per species, and $n = 250$ sites), accuracy was low, due to lack of information (fig. 5.5).

Species tree estimation was affected by genotyping errors at low read depths, especially at the high base-calling error rates (0.01 and 0.005). Indeed estimation accuracy was lower at read depths 8 than at read depth 3-5. This counter-intuitive result is due to the discrete nature of read depth: for example, at the base-calling error rate $e = 0.01$, the expected genotyping-calling error for heterozygotes is

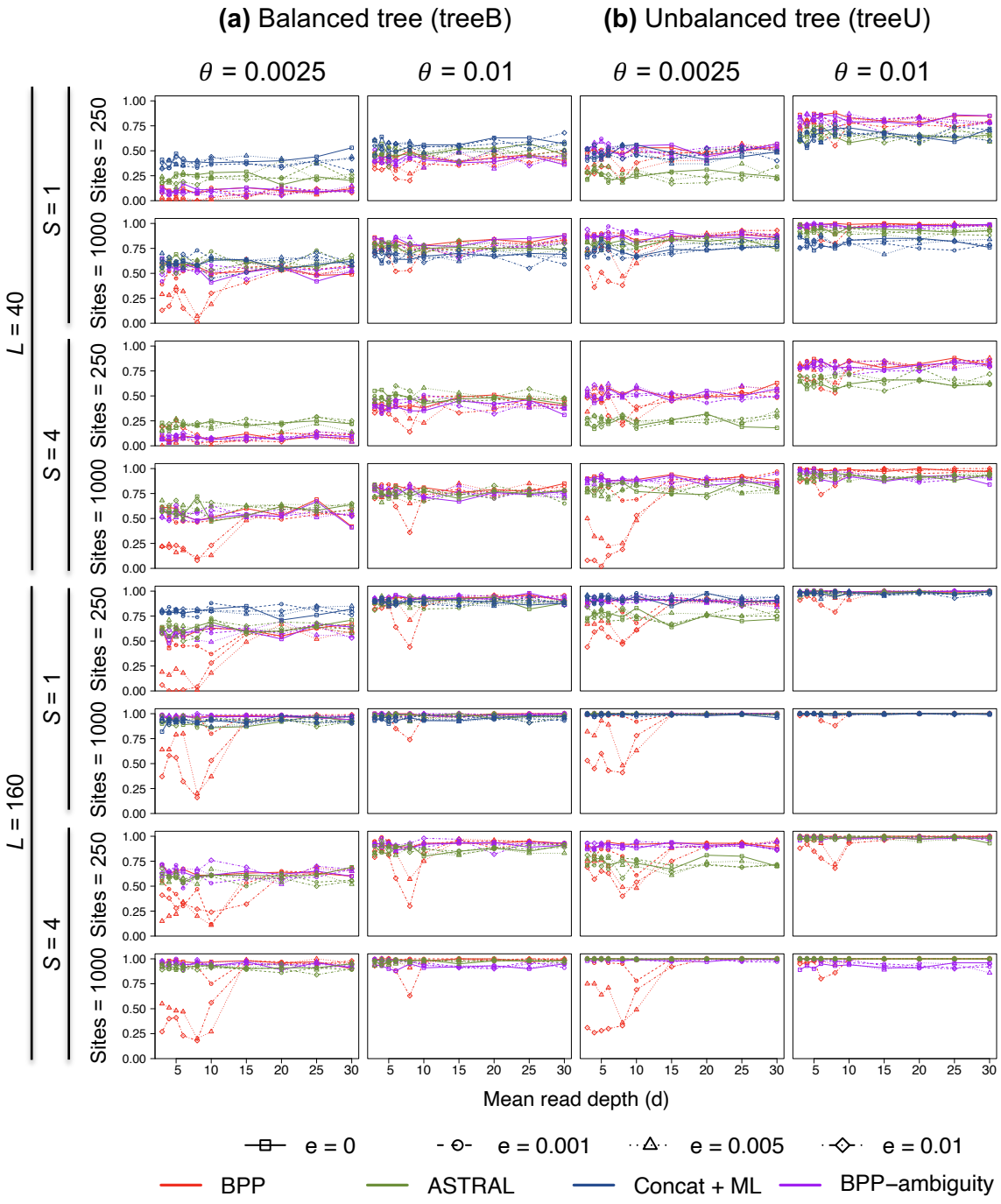

Figure 5.5: Accuracy of species tree estimation using BPP, BPP-ambiguity, ASTRAL and concatenation/ML. The results are shown separately for the four methods in figures S5.1&S5.2. Concatenation/ML is applied to the case of one (diploid) sequence per species ($S = 1$) only.

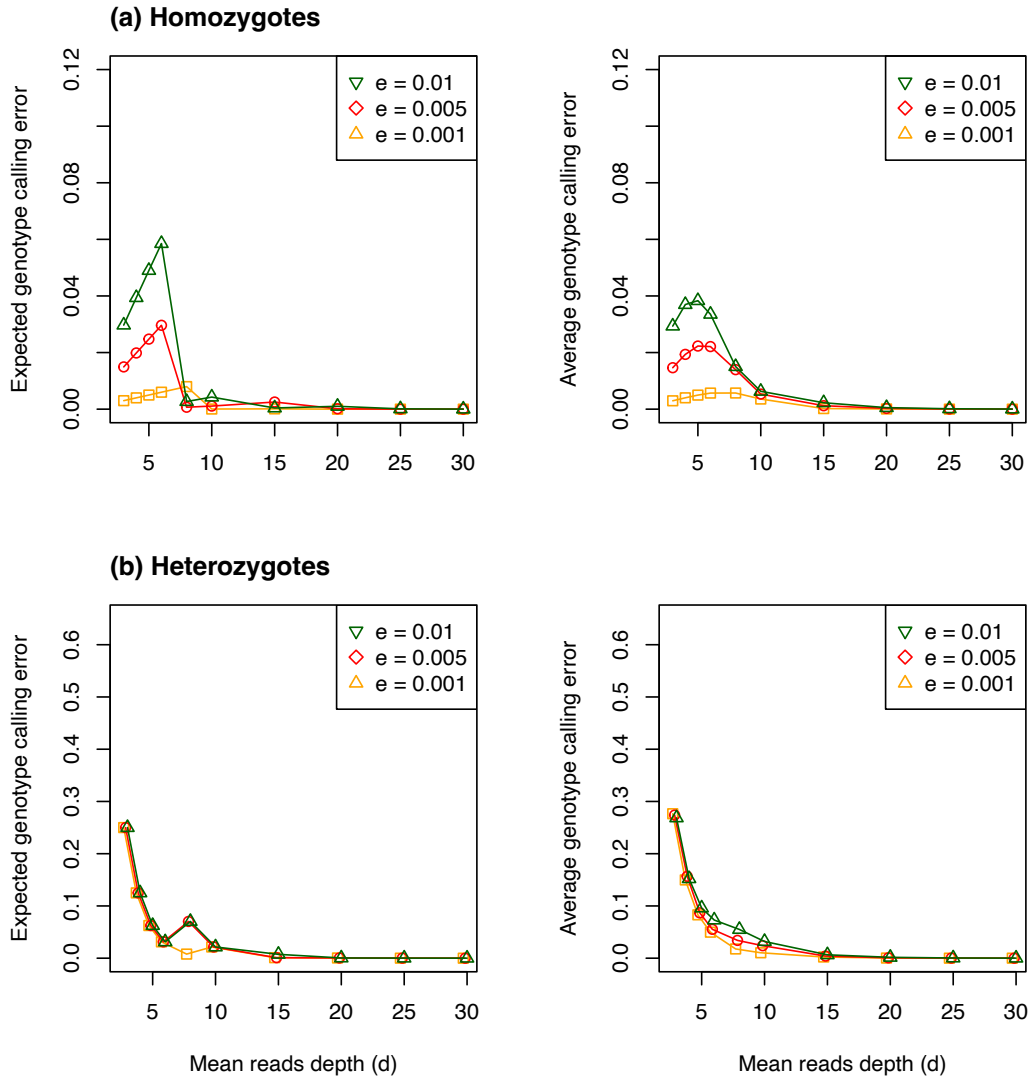

Figure 5.6: Expected and observed genotyping error rate for given base-calling error (ε) and read depth (d), as in Thawornwattana *et al.* (2018). In our simulation, we used the error rates $\varepsilon = 0$ (no errors), 0.001, 0.005, and 0.01. Note that when $\varepsilon = 0$ the genotyping error rate is 0 for homozygotes and $\left(\frac{1}{2}\right)^{d-1}$ for heterozygotes.

higher at read depth $d = 8$ than at $d = 5$ (fig. 5.6). When average read depth is 15X or higher, genotyping errors did not have an effect on species tree anymore in our simulation.

Accuracy of the probability of inferring the true tree reaches its peak at a mean read depth of 15X for $\theta = 0.0025$ and 10X for $\theta = 0.01$. This likely indicates that genotyping errors caused by base-calling inaccuracies are a major factor distorting tree inference in BPP when relatively low-depth sequencing data (< 5X) are used. However, this issue diminishes when sequencing depth exceeds 15X or 10X, depending on the divergence between species.

Higher mutation rate ($\theta = 0.01$ versus 0.0025) is noted to improve the accuracy of species tree

inference. Note that in our experiment, species split times (τ) are proportional to θ, so that the shape of the species tree does not change with θ, and θ mimics the use of different genomic regions with different mutation rates (e.g., neutral DNA versus conserved noncoding elements or exons).

Accuracy was higher for the unbalanced tree U (fig. 5.2) than for the balanced tree B. As pointed out by Huang *et al.* (2020), this is due to our choice of the internal branch lengths: in tree B, the three internal branch lengths have the lengths 0.1θ, 0.2θ, and 0.2θ, whereas in tree U, all three internal branches had the length 0.2θ. Coalescent is less likely to occur on branches of short length, and there is limited information on the local topology, making it prone to phylogenetic errors.

Next we consider the approach of treating called heterozygotes as ambiguities (fig. 5.5, BPP-ambiguity). This treats heterozygotes as missing data (for example a T/C heterozygote, which means 'both T and C', is treated as 'either T or C'), and is expected to use less information in the data, and to underestimate heterozygosity or θ. The approach reduced the impact of genotyping errors considerably. Also with this approach, the results were very similar at different base-calling error rates and at different average read depths.

We also analyzed the same data using ASTRAL and concatenation/ML to estimate the species tree. We include a sequence from a distant outgroup species (O) to root the tree (fig. 5.2), as these methods do not infer the root of the species tree. For ASTRAL analysis, we used RAxML to reconstruct the gene tree for each locus under the JC model and then used ASTRAL to generate the species tree. The RAxML analysis treats the diploid sequence with heterozygotes as a haploid sequence with ambiguities; for example a T/C heterozygote is treated as either T or C (Andermann *et al.*, 2019; Huang *et al.*, 2022a). The results are summarized in figure S5.2. Overall, ASTRAL and concatenation/ML appear to be robust to genotyping errors in the simulations here. Performance was nearly identical at different read depths and at different base-calling error rates. The two methods performed better than BPP and were similar to BPP-ambiguity.

In the case of no sequencing errors ($\varepsilon = 0$), BPP most often had better performance than ASTRAL and concatenation/ML. However, exceptions do exist. For example in the case of tree B, and lower mutation rate ($\theta = 0.0025$), concatenation/ML performed better than ASTRAL, which in turn was better than BPP. Such unusual cases are uncommon but are expected to occur (see Yang, 1996, 1998

for discussions).

5.2.3 Parameter estimation under the MSC-I model

We simulated data using the balanced and unbalanced trees (B and U) for five species of figure 5.3 with two gene-flow events and analysed the data using BPP to estimate parameters in the model. The true introgression probabilities $\varphi_{bc} = 0.3, \varphi_{de} = 0.2$.

First we discuss the results under the MSC-I model. The average posterior means and HPDs among the replicates are presented in figures 5.7 and S5.4–S5.6.

In the standard BPP analysis (BPP in fig. 5.7 & S5.4–S5.6), low depth ($< 15X$) causes bias to many parameters in presence of base-calling errors ($\varepsilon > 0$). Population sizes for modern species (θ_A, θ_B, θ_C, θ_D and θ_E) and speciation/introgression times (τ_R, τ_S, τ_T, τ_U and τ_b, τ_d under the MSC-I model) are overestimated when read depth is $<$ 10X, with the peak mainly occurring at 5X. The θs for ancestral populations are generally much more robust to the errors.

Generally, parameters showed greater bias relative to the true values at the lower mutation rate ($\theta = 0.0025$ vs. 0.01). In theory, parameters related to short branches are estimated with more uncertainty, and the shortest branch in the trees of figure 5.3 is $u - r$ in tree B and $u - t$ in tree U, with a length of 0.5θ. The mean estimates of θ_U in both trees basically have no error, while the HPD intervals are slightly wider compared to other population sizes. Likewise, introgression probabilities $\varphi_{b\to c}$ and $\varphi_{d\to e}$ are biased with low-depth data.

We also examined the power to detect gene flow using the Bayesian test (Ji *et al.*, 2023) (figs. 5.7 and S5.4–S5.6, $P_{b\to c}$ and $P_{d\to e}$). In the test, we used a null interval of (0, 0.001) and confirmed that use of (0, 0.01) and (0, 0.005) produced nearly identical results. At the high base-call error rate ($\varepsilon = 0.01$) and with $S = 4$ sequences per species, power may be affected by genotyping errors at low read depths ($d = 5$). However, overall the power to detect gene flow is high. When base-calling errors are incorporated and depth is low ($d < 10$ and $\varepsilon > 0$) , introgression probabilities under the MSC-I model are estimated with large intervals, resulting in a low rate of detecting gene flow using the Bayesian test of gene flow, and it is more problematic for shallower trees with θ = 0.0025 (fig. S5.4 & S5.6). Introgression from d to e is expected to be easier than b to c in both trees, despite

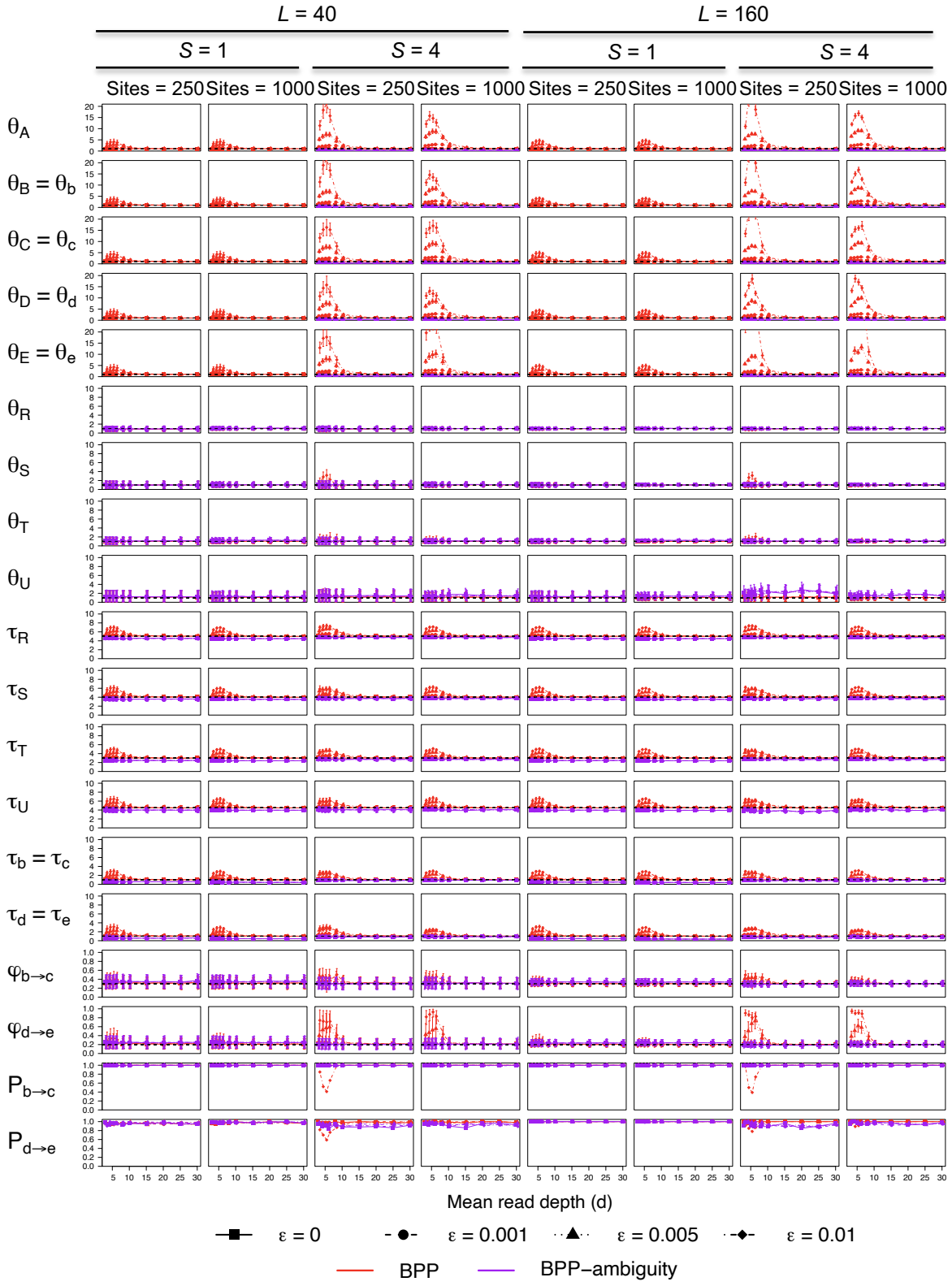

Figure 5.7: Average posterior means and 95% HPD CIs for parameters in data simulated and analysed under the MSC-I model of tree B (fig. 5.3a) with $\theta = 0.01$. Dashed lines indicate true parameter values (τ and θ are multiplied by 100). $P_{b\to c}$ is the power of the Bayesian test or the proportion of replicate datasets in which the Bayesian test inferred $b \to c$ introgression at the 1% level (with $B_{10} \geq 100$), using a null interval (0, 0.001).

of an exception in tree B with small amount of divergence ($\theta = 0.0025$) (fig. S5.4), where $P_{b\to c}$ is somewhat higher than $P_{d\to e}$ when depth is low. Both introgression events are detected in the test by $\sim 100\%$ when depth is 15X or higher except in the simulation with too limited data (L = 40, S = 1, Sites = 250).

We may ask the question whether a few samples sequenced at a higher depth are better than many samples sequenced at a lower depth. For example, the two scenarios $S = 1$ with $d = 20$ and $S = 4$ with $d = 5$ may involve similar amounts of sequencing effort or cost. The answer to this question is clear-cut: a few high-depth samples are much better than many low-depth samples. Indeed at $d = 5$, use of $S = 4$ samples exacerbates the bias and is worse than having one sample ($S = 1$ at $d = 5$), not to mention one sampled sequenced at great depth ($S = 1$ at $d = 20$).

Treating heterozygotes as ambiguities (missing data) (fig. 5.7 & S5.4–S5.6, BPP-ambiguity) reduced the bias in parameter estimation caused by genotyping errors at low read depth, although population sizes at tips (θ_A, θ_B, θ_C, θ_D and θ_E) are all underestimated. With heterozygotes treated as ambiguities, there is little difference between $S = 1$ and $S = 4$.

We note that for this set of simulation, the tree shape (B versus U) had little impact. The number of sequences ($S = 1$ or 4) has minimal impact, with the bias at low read depths tend to be more serious when more sequences per species are included.

5.2.4 Parameter estimation under the MSC-M model

Under the migration (MSC-M) model, the population migration rates used were $M_{bc} = 0.3, M_{de} = 0.2$. Here the population migration rate $M_{xy} = m_{xy}N_y$ is the expected number of $x \to y$ migrants, where m_{xy} is the proportion in the recipient population y of immigrants from x, and where N_y is the effective population size of y. The results for simulation under the MSC-M model are shown in figure 5.8, and S5.7–S5.9.

The impact of low depth on parameters estimated under the MSC-M model is similar to that in the MSC-I model. Most of parameters are overestimated due to the impact of genotyping errors, while there are few instances of underestimation under the MSC-M model, such as τ_U in tree B (fig 5.8 & S5.7). The underestimation of τ_U or the overestimation of spanning time for migration d to e is

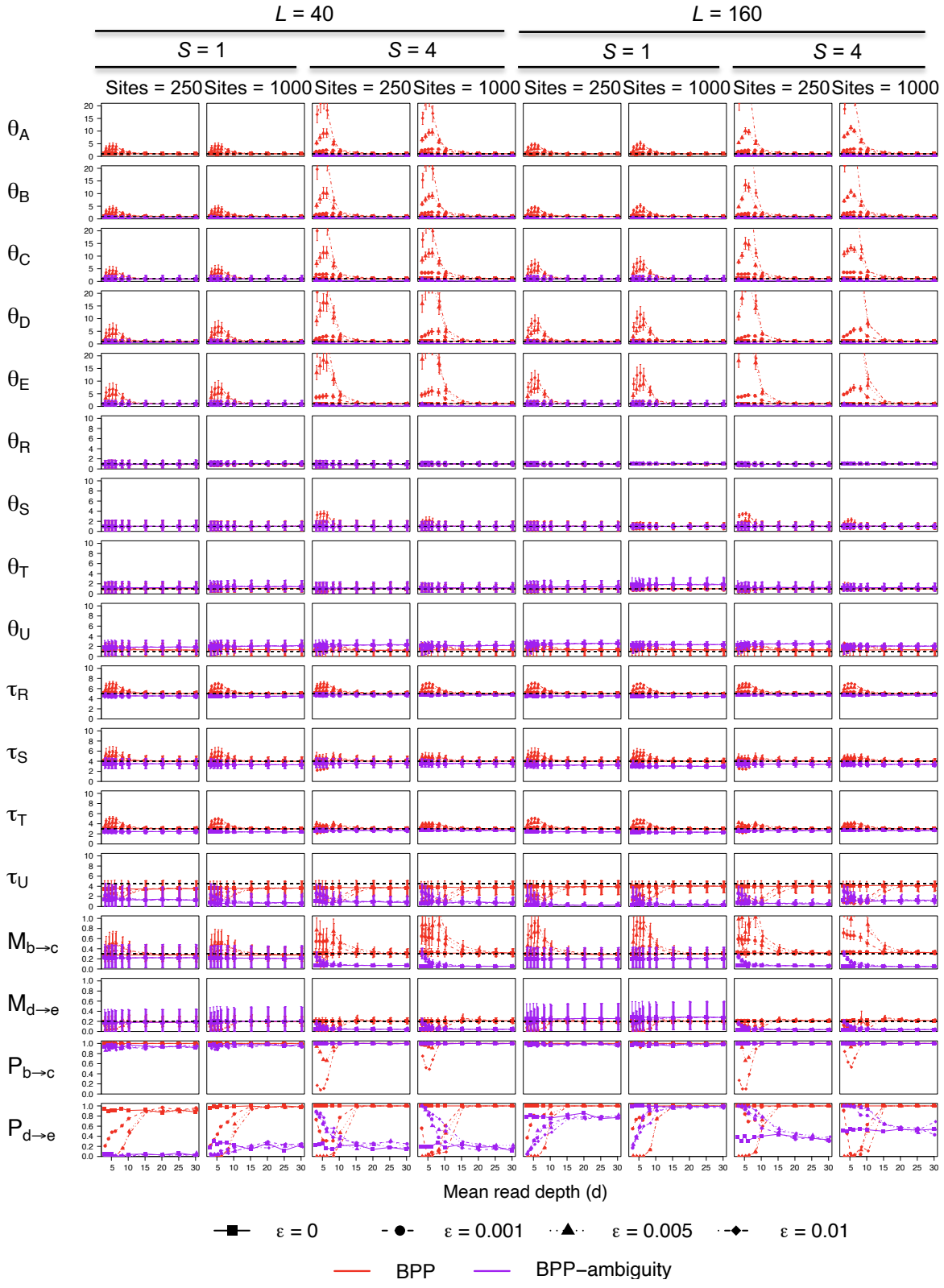

Figure 5.8: Average posterior means and 95% HPDs for parameters under the MSC-M model of tree B (fig. 5.3a) with $\theta = 0.01$. $P_{b\to c}$ and $P_{d\to e}$ is the power of the Bayesian test of gene flow (at the cutoff $B_{10} \geq 100$). See legend to figure 5.7.

essentially correlated with the underestimation in migration rate $M_{d\to e}$. Actually, in the absence of sequencing errors ($\varepsilon = 0$), the time τ_U is not very well inferred and underestimated in the cases where simulated datasets contain a limited number of loci ($L = 40$)

Some parameters under the MSC-M model appear to be more influenced by low depth. The migration rates $M_{b\to c}$ and $M_{d\to e}$ have larger bias than the introgression probabilities φs in the MSC-I model above. The estimates under the MSC-M model are close to the truths with acceptable error at depth of 15X, and 20X appears to be the safe choice to ensure accurate inference of all parameters in specific cases — for example, $M_{d\to e}$ under tree B with $\theta = 0.0025$ (fig. S5.7). Overall, the MSC-M model turns out to be more demanding and expects high-depth data of at least 20X.

As shown for BPP-ambiguity in figure 5.8, and S5.7–S5.9, treating heterozygous sites as ambiguities is also useful for mitigating the impact of genotyping errors in parameter estimation under the MSC-M model. Population sizes for ancestral populations and speciation times (except τ_U) are estimated with sufficient accuracy and almost no bias across depths 3X to 10X. For the modern population sizes, it is the same as in the MSC-I model where they are underestimated because of the use of unphased sequences.

In our simulation, there is very little improvement for involving more loci ($L = 160$) and more sites per sequence (Sites = 1000), and the use of larger datasets does not always help reduce the bias. Notably, when phasing is disabled in BPP and there are multiple sequences per species (fig. 5.8 BPP-ambiguity, $S = 4$), migration rates estimates are biased, and some of them are close to 0. This underestimation is probably caused by analysing data with approach BPP-ambiguity as the impact of genotyping errors become negligible when depth exceeds 20X.

Regarding the impact on the Bayesian test of gene flow, the power of the test is compromised under the migration model due to the wide HPD intervals of M estimates using low-depth data. The power of the test reaches 100% when depth is 15X or 20X.

Parameters tend to have smaller bias if heterozygotes are regarded as ambiguities, although this approach leads to underestimation in migration rates when $S = 4$ sequences are present for each species. The Bayesian test of gene flow is largely robust to the bias and detects evidence of gene flow in most replicate datasets. However, there are cases where the test yields low power. In figure

5.8, with $S = 4$, the high power $P_{b\to c}$ of BPP-ambiguity is a result of the overestimation in $M_{b\to c}$ when depth is $< 10X$, which then gradually decreases and stabilizes at a relative low level as depth increases. Thus, we suggest that the analytic strategy of BPP-ambiguity should be used with caution under this model.

5.3 Discussion

5.3.1 Limitations of our simulation of read depth

In this chapter we develop a model for simulating read depths for sites in a sequence. The model accounts for the strong correlation between read depths at two adjacent sites. Here we discuss the limitations of the simulation model and of our simulation. First the model of read depths is Markovian in that the read depth at the next site depends on the read depth at the current site, but not on read depths at the previous sites. This assumption is clearly unrealistic. It should be simple to incorporate high-order dependence. However for the purpose of our study, which is to assess the impact of genotyping errors on inference under the MSC, we suggest that the assumption is unimportant. As the read depths at adjacent sites is very strong, the difference between the models is minor.

Our simulation model has not accounted for mapping or alignment errors. In particular, if genomes are sequenced from different species and if the reference genome is far away, mapping errors may be considerable. When selecting genomic regions for generating multi-locus data, we typically avoid regions such as simple repeats and transposable elements, which effectively minimizes the alignment errors in the data (Gronau *et al.*, 2011). Another issue is that in our simulation, we assumed that genotype calling is based on reads for each sample. In the case of multiple samples from one species, use of multiple samples to call genotypes is known to reduce genotyping errors (Poplin *et al.*, 2017). Methods for calling genotypes using multiple-sample read data from several species are yet to be developed.

Despite those limitations, we suggest that our simulation provides useful guidelines for genome-sequencing projects for inferring species phylogenies accounting for the coalescent process, and for inferring interspecific gene flow.

5.3.2 Approaches to dealing with genotyping errors

The inference devices in BPP are affected by the genotyping errors in low-depth data, generating biased parameter estimates. Currently, there are several approaches to dealing with genotyping errors at different levels.

The first approach is to develop methods that work in data with genotyping errors. Zhang and Nielsen (2025) developed a method called WASTER for inferring species trees using data of low coverage, based on the method called CASTER from Zhang *et al.* (2025). These are heuristic methods that use genome-wide site pattern counts and ignore information in the variation of genealogical histories across the genome. The methods are aimed to estimate the species tree topology only, and do not provide estimates of population demographic parameters such as population sizes, species split times, and rates of gene flow between species.

In the likelihood framework, Gronau *et al.* (2011) developed a Bayesian method, called BSNP, that infers genotypes at each site using information in the aligned bases, including base-call quality scores, and mapping quality scores produced by BWA (Li and Durbin, 2009). The aim is to infer the correct, unbiased genotype at each position and prevent genotype calling from being driven by reads from low-coverage genomes. To some extent, the method reduces the number of erroneous genotype calls in genomes. This method may still struggle to handle tricky cases where all samples have uniformly low sequencing depth. One possible solutions is to accommodate genotyping uncertainty in phylogenomic and population genetic analyses (Korneliussen *et al.*, 2014).

From our simulation, we suggested the approach of treating heterozygotes as ambiguities is useful for reducing the biases. Miscalling of heterozygous sites as homozygotes are expected to be more damaging than calling homozygotes into heterozygotes, as the former type of errors potentially creates chimeric sequences (fig. 5.1). When treated as ambiguities, heterozygotes miscalled homozygotes merely result in some information loss but do not produce wrong phase resolutions, reducing the impact of the errors and making the inference in BPP more accurate.

5.4 Supplemental Information

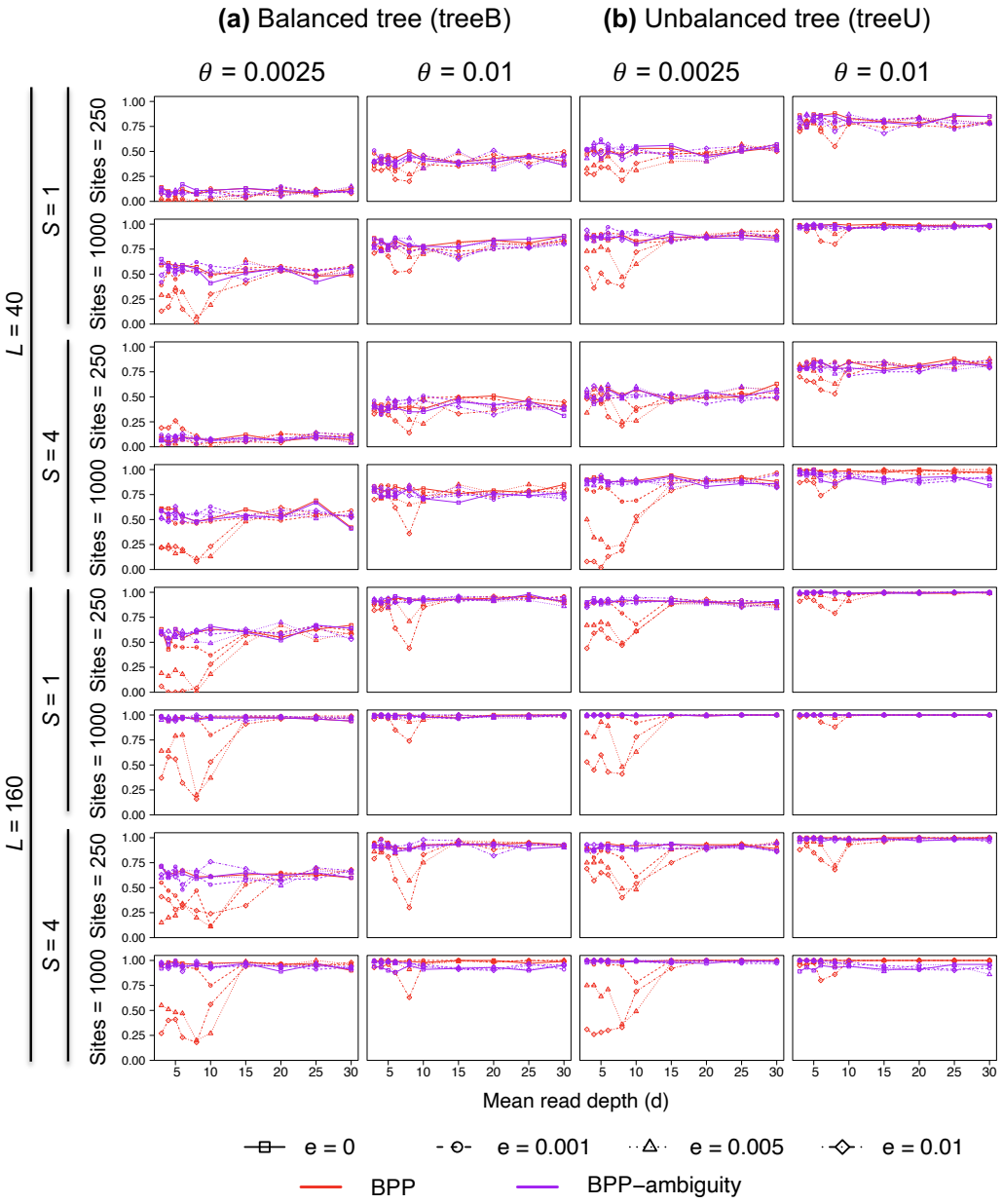

Figure S5.1: Accuracy of BPP species tree estimation at different mean read depths ($\bar{d}$) and base-calling error rates (ε), measured by the proportion of replicates in which the inferred species tree by BPP (the MAP tree) is correct.

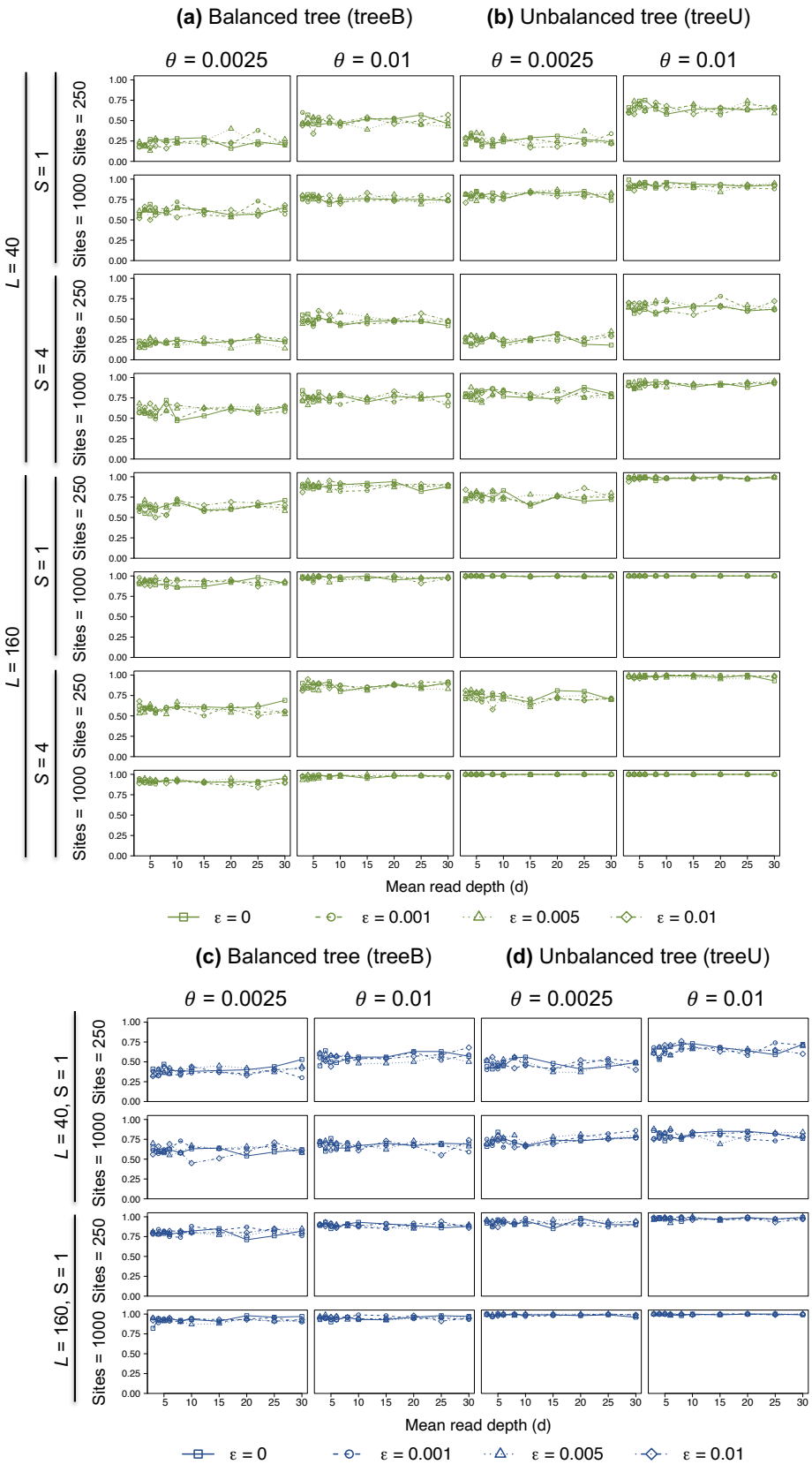

Figure S5.2: Accuracy of species tree estimation using (**a**, **b**) ASTRAL and (**c**, **d**) concatenation/ML at different mean read depths ($\bar{d}$) and base-calling error rates (ε). Data are the same as those of figure S5.1 except that an outgroup (O) is also included to root the tree. Concatenation/ML is applied to data of one (diploid) sequence per species ($S = 1$) only.

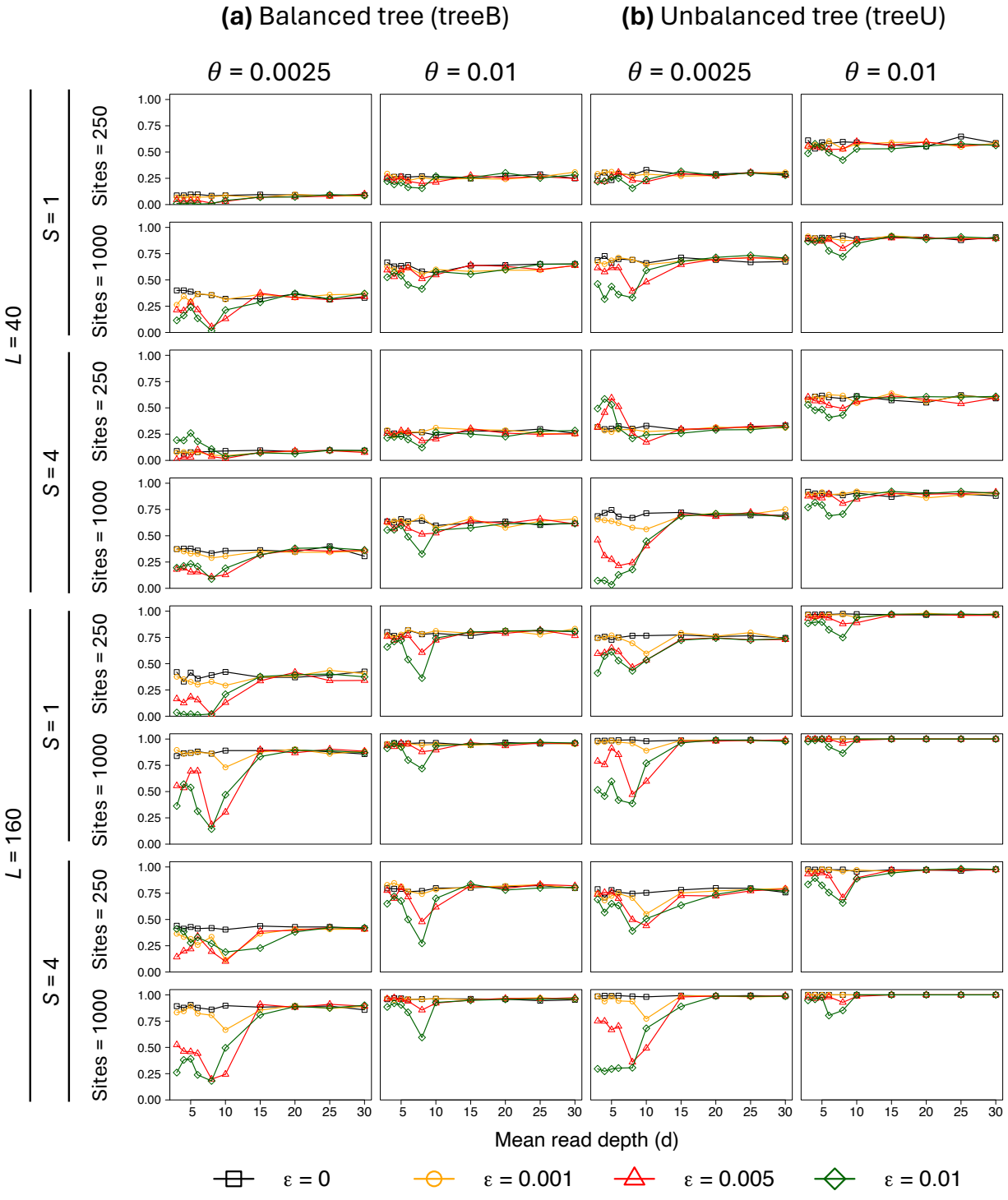

Figure S5.3: Average posterior probabilities for the correct species tree in BPP species tree estimation using simulated data at different mean read depths ($\bar{d}$) and base-calling error rates (ε). The true species trees are trees B and U of figure 5.2. The results are summarized using the same runs as BPP in fig. 5.5.

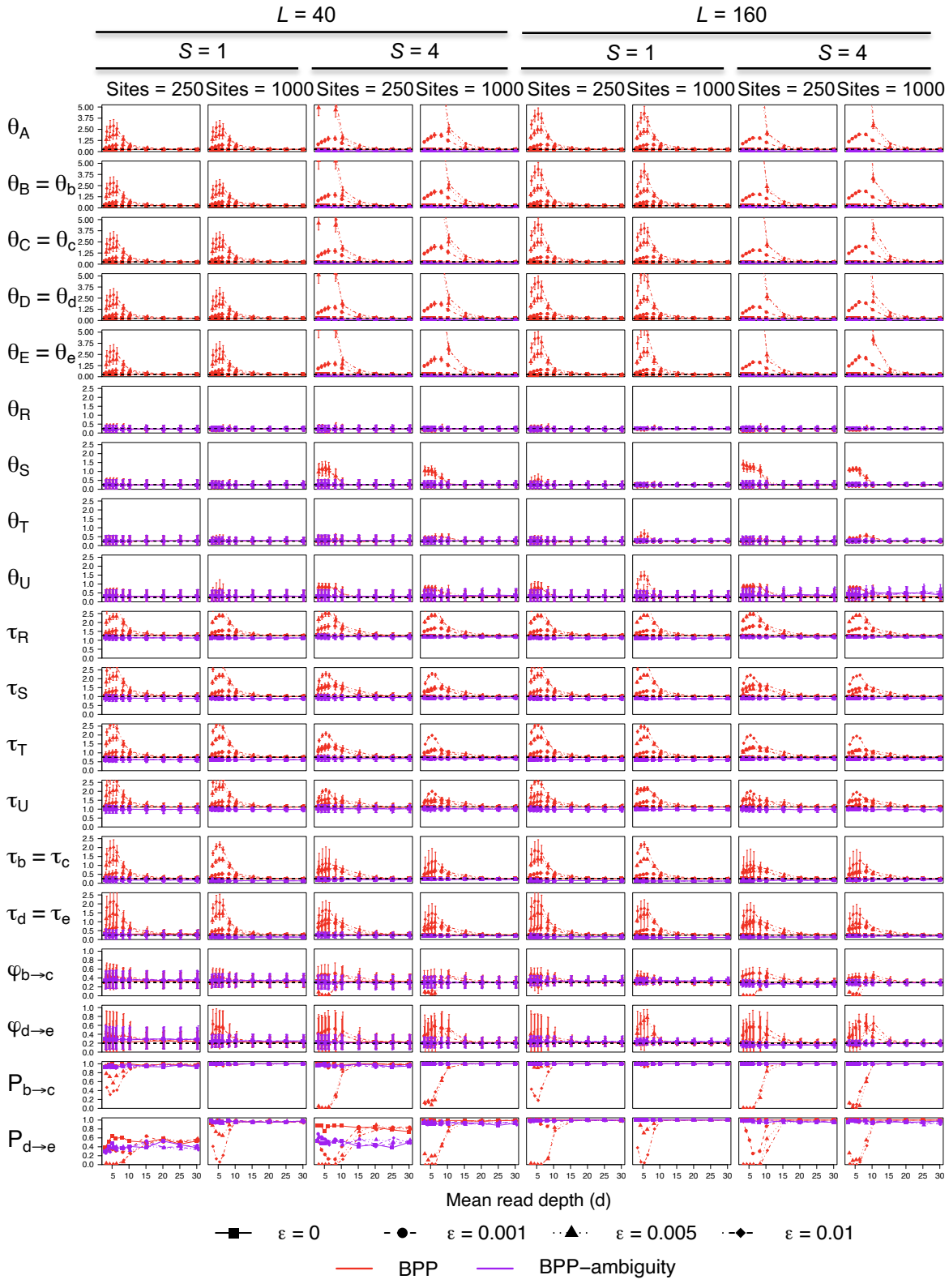

Figure S5.4: Average posterior means and 95% HPDs for parameters under the MSC-I model of tree B (fig 5.3a) with $\theta = 0.0025$. See legend to figure 5.7.

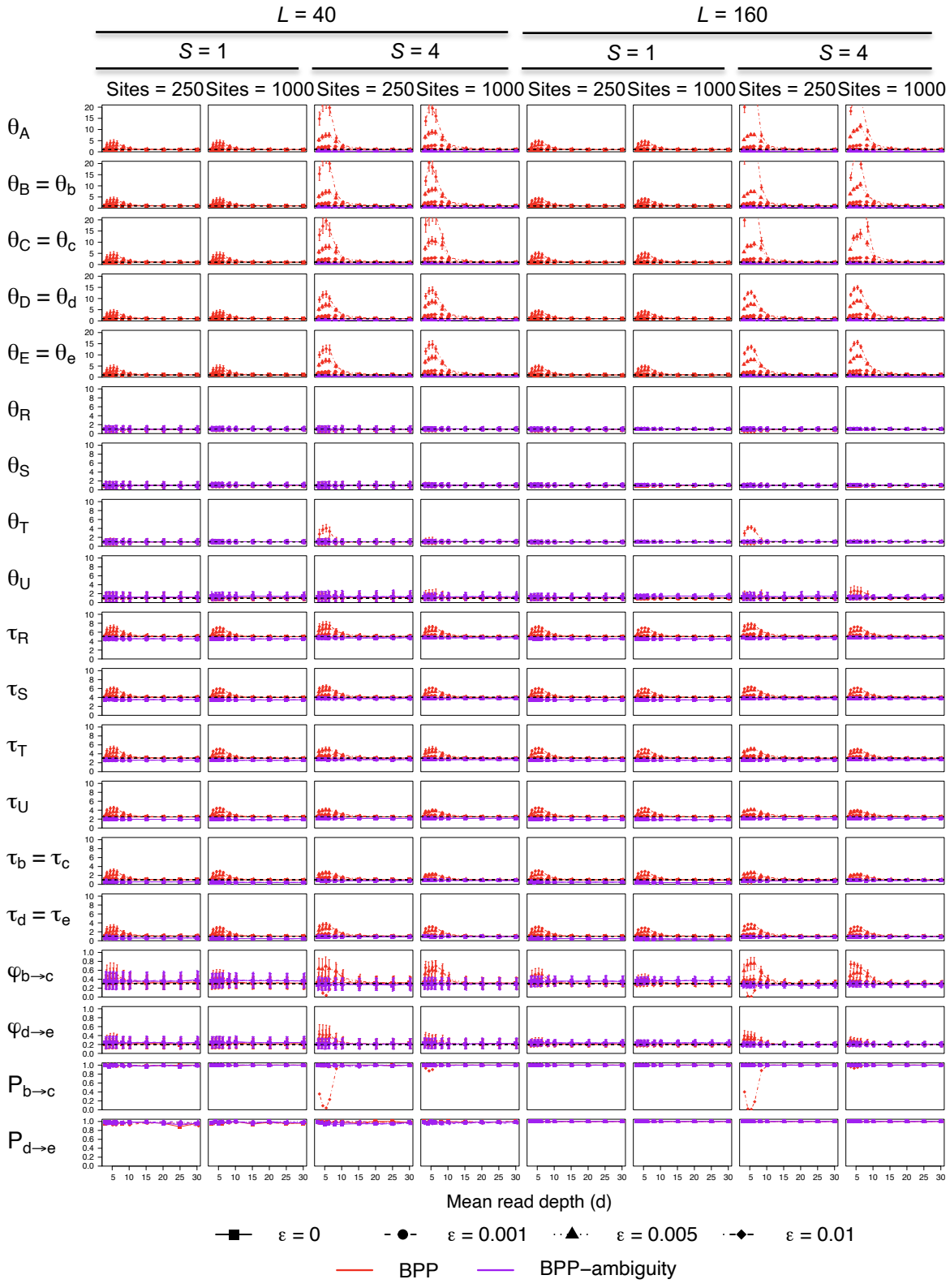

Figure S5.5: Average posterior means and 95% HPDs for parameters under the MSC-I model of tree U (fig. 5.3b) with $\theta = 0.01$. See caption to figure 5.7.

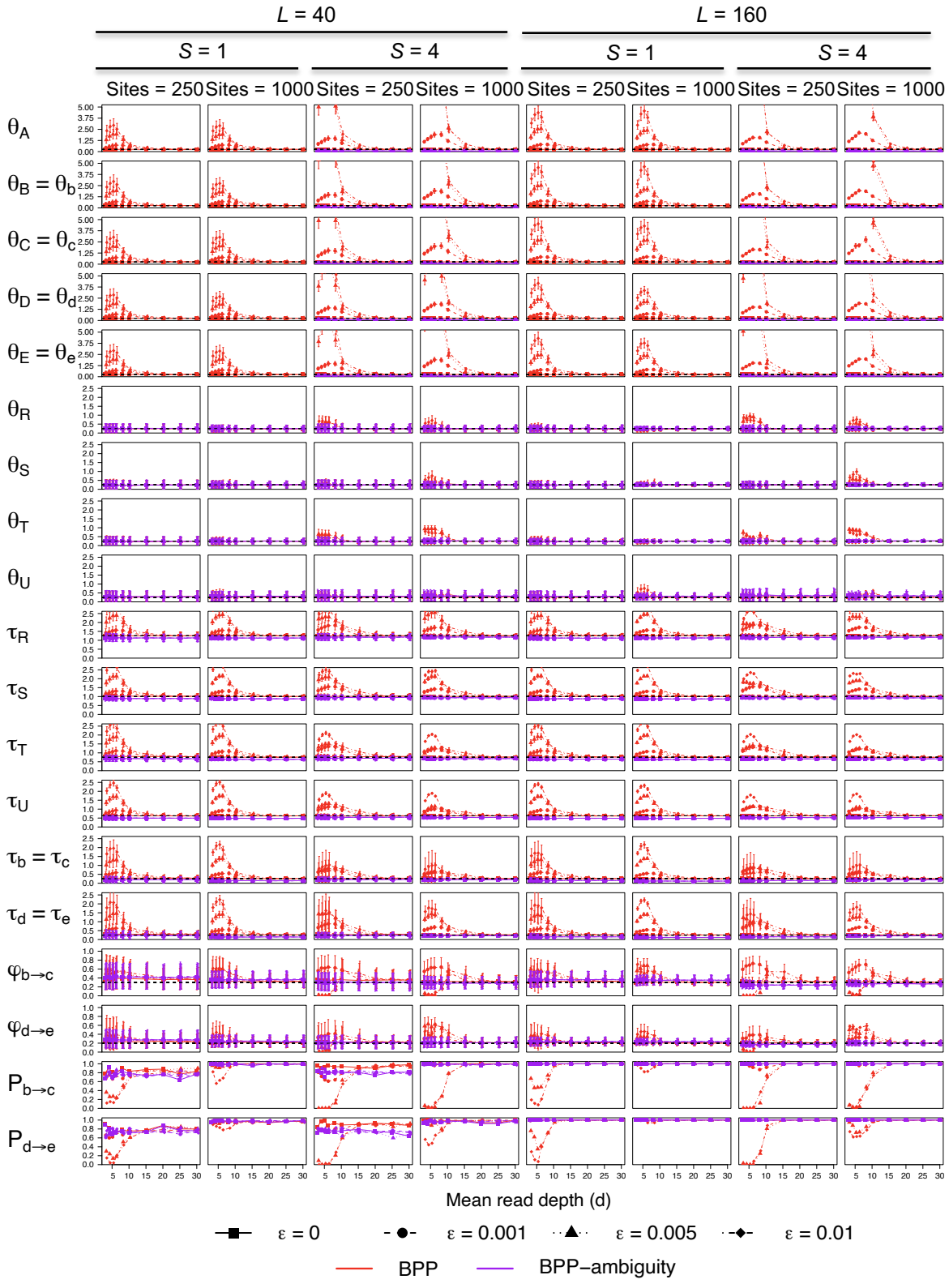

Figure S5.6: Average posterior means and 95% HPDs for parameters under the MSC-I model of tree U (fig. 5.3b) with $\theta = 0.0025$. See caption to figure 5.7.

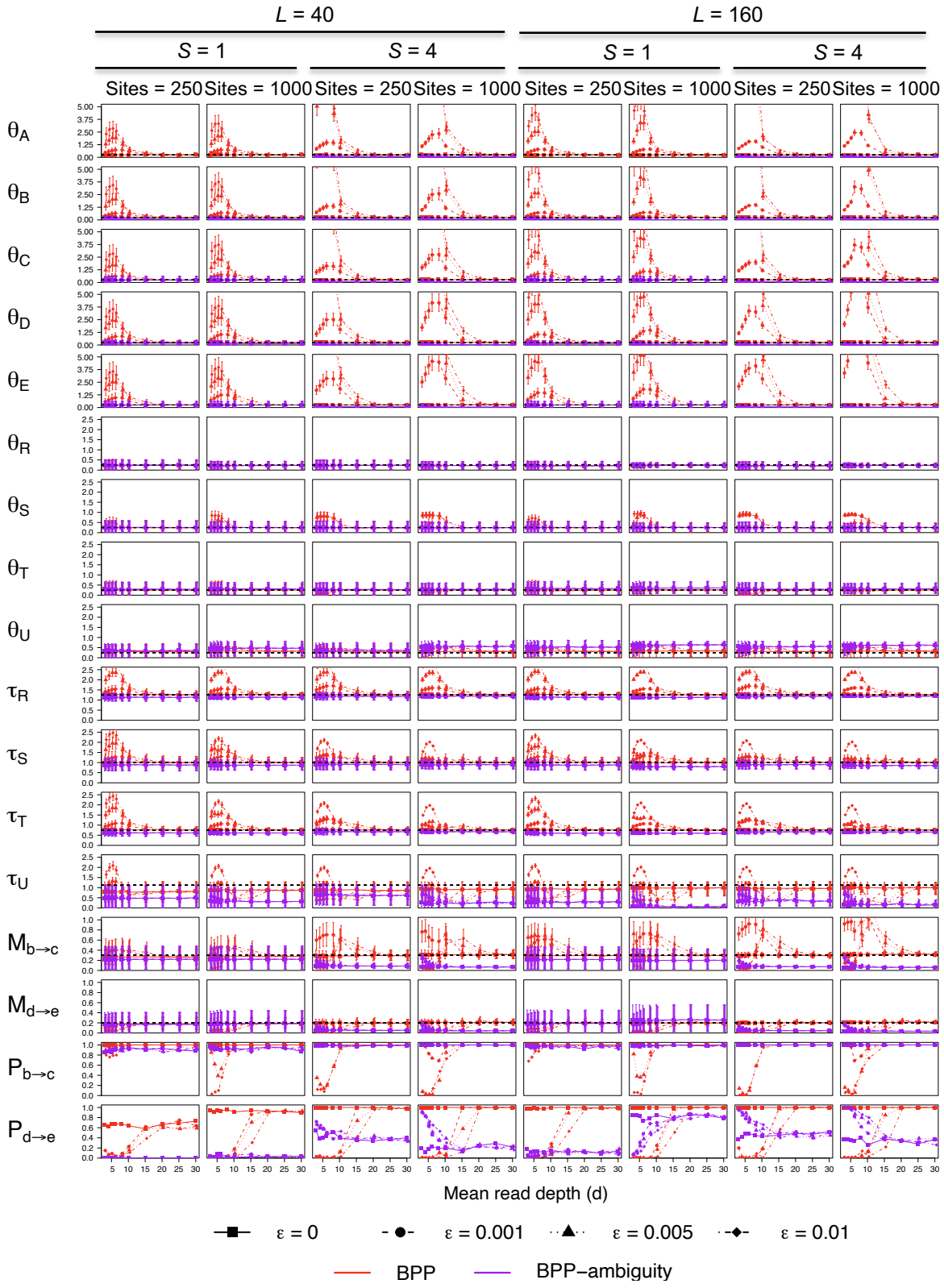

Figure S5.7: Average posterior means and 95% HPDs for parameters under the MSC-M model of tree B (fig. 5.3a) with $\theta = 0.0025$. See caption to figure 5.7.

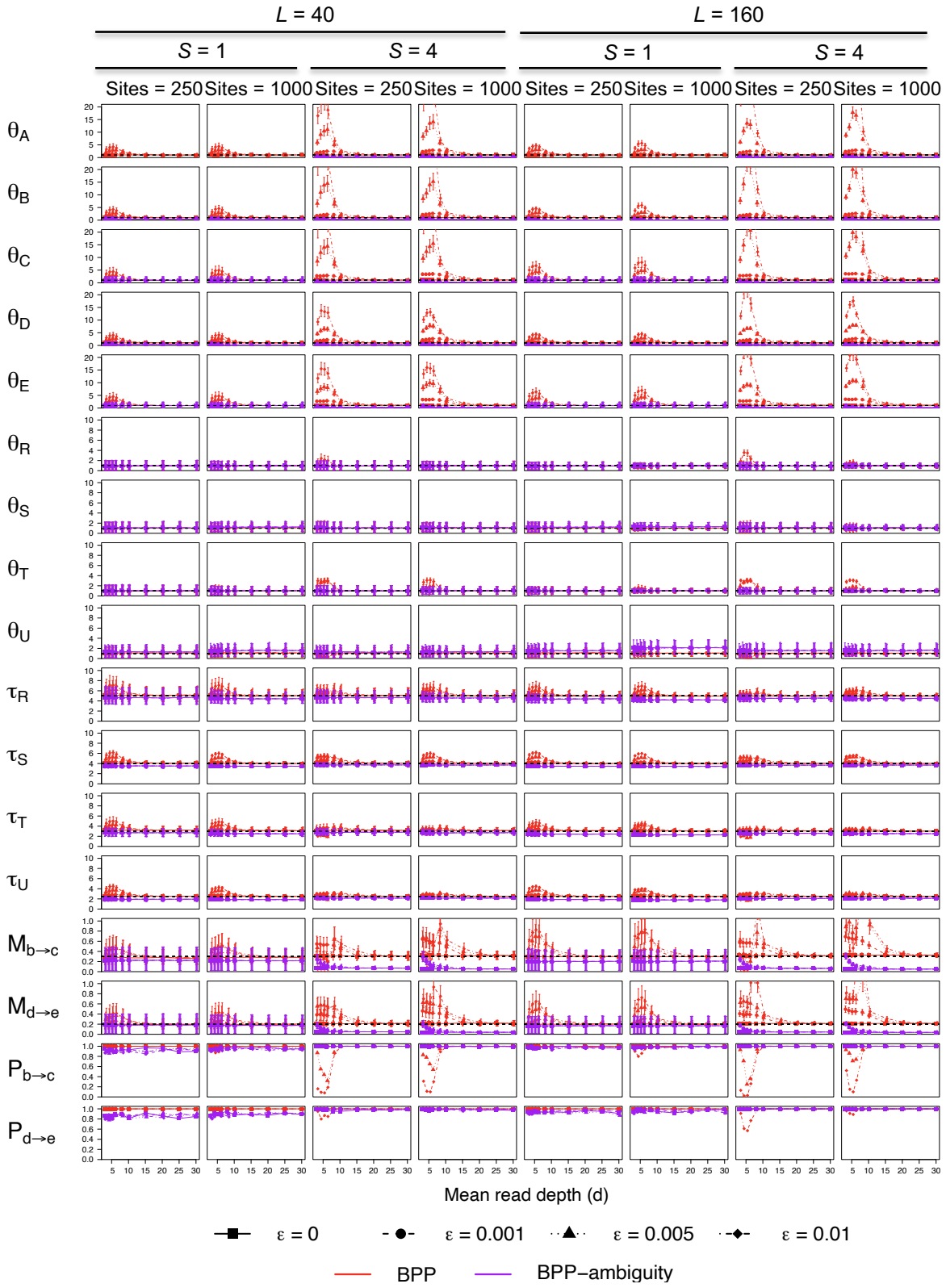

Figure S5.8: Average posterior means and 95% HPDs for parameters under the MSC-M model of tree U (fig. 5.3b) with $\theta = 0.01$. See caption to figure 5.7.

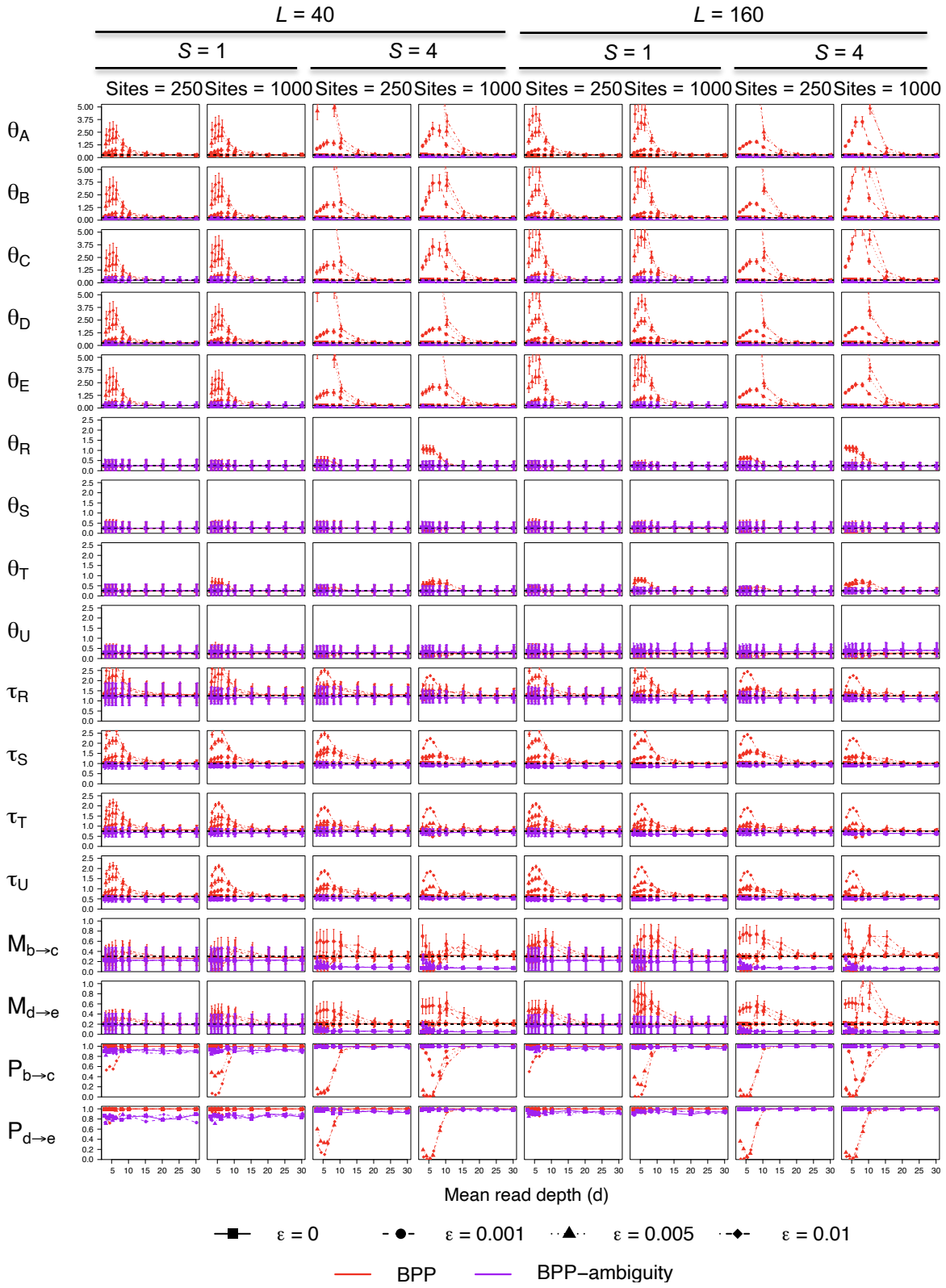

Figure S5.9: Average posterior means and 95% HPDs for parameters under the MSC-M model of tree U (fig. 5.3b) with $\theta = 0.0025$. See caption to figure 5.7.

Table S5.1: The expected and observed probabilities of correlated read depth between adjacent sites.

	Chimpanzee genome (20.26X)		Rabbit genome (4.21X)	
	P_1	P_2	P_1	P_2
Expected				
$p = 0$	0.2275	0.6143	0.5549	0.9668
0.1	0.2387	0.6377	0.5745	0.9751
0.2	0.2493	0.6594	0.5927	0.9813
0.3	0.2595	0.6795	0.6099	0.9859
0.4	0.2693	0.6981	0.6264	0.9893
0.5	0.2789	0.7156	0.6428	0.9918
0.6	0.2881	0.7319	0.6601	0.9936
0.7	0.2971	0.7472	0.6801	0.9949
0.8	0.3061	0.7617	0.7044	0.9955
0.9	0.3156	0.7762	0.7415	0.9951
Calculated	0.5729	0.9114	0.9410	0.9956

Note.– The expected values are calculated under Model 1 using $p = 0$ to 0.9, while the line at bottom displays the proportions obtained from the real data. $P_1 = P(d_{\text{next}} = d_{\text{current}}) = \sum_i \pi_i P_{ii} = \sum_i f_{ii}$ is the probability that the next site has the same read depth as the current site, and $P_2 = P(|d_{\text{next}} - d_{\text{current}}| \leq 1) = \sum_i \pi_i (\sum_{j=i-1}^{i+1} P_{ij}) = \sum_i \sum_{j=i-1}^{i+1} f_{ij}$, represents the probability that the next site has a read depth that differs by at most 1 from the current site. For both samples, given the depth $d_{current}$ at the current site, more than 90% sites at the next position along the genome have a depth within the interval $[d_{current} - 1, d_{current} + 1]$.

Summary

Gene flow is now recognized as a common feature of species divergence in many taxa, with important evolutionary consequences for local adaptation and species diversification. In the first three chapters, we demonstrate the superior power of full-likelihood approaches under the multispecies coalescent (MSC) framework for inferring gene flow. We also develop a Bayesian test of introgression, which computes Bayes factors using the Savage–Dickey density ratio from MCMC samples. This test can be used to assess the significance of individual gene flow events.

In Chapters 2 and 3, we adopt a stepwise approach to construct models of introgression or migration, while in Chapter 4, we construct a joint migration model by summarizing evidence from triplet analyses. Across these studies, we identify substantial gene flow among closely related species and evaluate the performance and limitations of commonly used summary statistics. In Chapter 5, we investigate the impact of sequencing depth on inference and show that treating heterozygotes as ambiguities in low-coverage data can effectively mitigate the bias caused by genotyping errors. Together, the thesis highlights an urgent need to apply likelihood methods for inferring gene flow using genomic sequence data and to improve the statistical properties of summary methods.

Stepwise construction of model of gene flow Inferring species trees with gene flow is both biologically crucial and computationally challenging. Bayesian methods are struggling with exploring the space of possible introgression or migration models, even for datasets with as few as 100 loci (Jiao *et al.*, 2021). The stepwise model construction offers a statistically principled and computationally feasible alternative.

The stepwise approach works on a stable species tree that is not misled by the gene flow. Despite the misspecification, the species tree can often be accurately inferred using methods ignoring gene flow, such as the MSC model implemented in BPP. For scenarios with excessive introgression and rapid speciation, the species tree can be highly uncertain, and the posterior may involve multiple competing species trees. It is still possible to identify the true species tree from genomic regions that are less affected by gene flow. For example, when migration results from sex-biased dispersal, the sex chromosomes or mitochondrial genomes specific to the opposite sex may remain unaffected. As a re-

sult, individuals from each population/species tend to form monophyletic groups on the corresponding gene trees. In Chapters 4, we identify the population phylogeny for chimpanzees and bonobos using mitochondrial data. Additionally, regions of low recombination tend to exhibit reduced introgression, with certain parts of the genome being most resistant to gene flow. In *Heliconius* butterflies, the Z chromosome shows no evidence of recent gene flow, while autosomes display widespread admixture (Thawornwattana *et al.*, 2022, 2023b).

In our stepwise framework, gene flow events are iteratively added to the binary species tree, usually in order of decreasing statistical support (e.g., estimated introgression probabilities). After each addition, the Bayesian test of gene flow is applied to determine whether the events in the model are significantly supported given the data; non-significant events are excluded from the joint model. Alternatively, it may begin with a saturated or nearly saturated model that includes all or most of gene flow events. The non-existent events are expected to be estimated with probabilities close to zero and drop out. These two strategies reflect alternative heuristics for navigating model space, and in theory, both are able to reach the true model. Although starting with a saturated model may require fewer steps to reach the final model, it involves fitting complex, parameter-rich models early on. In contrast, stepwise addition is often more computationally tractable, especially for large phylogenies.

When dealing with phylogenies involving 5 or more species/populations, the number of possible gene flow edges in the species tree becomes prohibitively large. A common workaround is to first identify introgression events among extant species, and then translate those signals into events involving ancestral lineages. For example, if gene flow is detected from species A to C and from B to C, this may be interpreted as introgression from their ancestor AB to C. Another promising and scalable approach is the divide-and-conquer strategy, which was originally developed for estimating species trees (Molloy and Warnow, 2019). Similarly, inferences of gene flow can be performed on subsets of species (e.g., triplets) and the resulting evidence of gene flow is merged parsimoniously to obtain a full model.

Despite the practical advantages, the stepwise approach has limitations. First, it does not perform an exhaustive search over model space and may miss some gene flow events. Furthermore, it assumes the availability of a well-supported species tree, which may not be feasible in cases with extremely

rampant introgression. Lastly, while the approach itself is applicable to datasets of any size, analysing huge datasets can still demand substantial computational resources. In such cases subsampling of loci/sequences may be employed to downsize the data.

Challenges and outlook The MSC-I and MSC-M typically require thousands of loci to obtain reliable estimates, particularly for models of many parameters, while performing model selection through introgression or migration models remains a major challenge when working with datasets at this size. Even with a fully specified model, parameter estimation based on full-likelihood methods can still be computationally demanding with more than 10 species involved.

Summary methods are computationally fast and scale well to large phylogenies. Despite their heuristic nature and the associated information loss, they remain useful especially in exploratory analyses to rapidly identify candidate introgression events over a species tree. Efforts have been made to integrate the computational efficiency of summary methods with the statistical rigour of full-likelihood inference. One such framework D-BPP (Yang *et al.*, 2025) unifies the summary method D-STATISTIC (Durand *et al.*, 2011b) and the full likelihood method BPP MSC-I (Flouri *et al.*, 2020), and it is designed to resolve complex introgression scenarios by evaluating competing models through Bayesian model comparison. Moreover, it should be possible to improve the power of summary methods by making more efficient use of multilocus information, including gene tree topologies, branch lengths, and variation in gene trees across loci (Jiao *et al.*, 2021).

Another direction is to generalize the MSC framework to incorporate other important biological processes such as recombination and natural selection.

Phylogenomic analyses under the MSC models assume independent gene trees across loci and a single gene tree per locus, implying free recombination among loci and no recombination within each locus. The estimation of gene flow may benefit from making use of information in linkage disequilibrium between adjacent genomic segments. Instead of assuming independence across loci, methods including PHYLONET-HMM (Liu *et al.*, 2014) and DICAL-ADMIX (Steinrücken *et al.*, 2018) use a hidden Markov model (HMM) to approximate gene tree correlations along the genome. These methods, typically developed for two or three taxa, can be used to estimate introgression and also identify introgressed genomic regions.

Selection is another factor that has not yet been accommodated into the MSC framework. Previous simulation studies found that species tree estimation tends to be robust to background selection and positive selection (Shi and Yang, 2018; Thawornwattana *et al.*, 2018, 2022). However, more recent work indicates that ignoring selection can bias the detection of introgression using full-likelihood methods (Smith and Hahn, 2024). There are methods that allow for variable θ among loci to reflect the reduced mutation rate in background selection. The incorporation of fitness effects in more complex scenarios of selection (e.g., balancing selection and directional selection) may rely on the use of machine-learning approaches (Mo and Siepel, 2023; Schrider and Kern, 2018).

When read depth is a concern in the data, such as in ancient DNA or other degraded samples, genotyping errors can lead to distorted inference. The approach of treating heterozygotes as missing data can mitigate the impact of genotyping errors, especially for analysis under the MSC-I model. It may be useful to develop multi-sample genotype-calling methods under the MSC model to improve genotyping quality at low read depths. If the species are closely related, even multi-sample genotype-calling procedures developed for population data (from one species) may improve genotyping quality. If low-depth sequence data are common, it may be worthwhile to implement probabilistic models to accommodate sequencing and genotyping errors in genome sequences at low depths.

Ultimately, advances in methods for reliably detecting and quantifying gene flow will enhance our understanding of its role across different evolutionary timescales. In shallow phylogenies, accurate inference of gene flow among closely related species will illuminate key biological questions including how often species diverge with gene flow between closely related species and how species can remain distinct in spite of gene flow (Mallet *et al.*, 2016). Studying introgression provides critical insights into the nature of species. Over deeper timescales, introgression contributes to biodiversity by creating novel genetic combinations that facilitate ecological adaptation and species divergence (Abbott *et al.*, 2013; Marques *et al.*, 2019). Improved methods will enable more accurate assessment of its prevalence over an extended timeline and refine our understanding of its macroevolutionary impact on species diversification and trait evolution. As we face a world of continuously reducing biodiversity, learning from past examples may improve our ability to anticipate responses to environmental change.

References

Abbott, R., Albach, D., Ansell, S., Arntzen, J. W., Baird, S. J. E., Bierne, N., Boughman, J., Brelsford, A., Buerkle, C. A., Buggs, R., Butlin, R. K., Dieckmann, U., Eroukhmanoff, F., Grill, A., Cahan, S. H., Hermansen, J. S., Hewitt, G., Hudson, A. G., Jiggins, C., Jones, J., Keller, B., Marczewski, T., Mallet, J., Martinez-rodriguez, P., Möst, M., Mullen, S., Nichols, R., Nolte, A. W., Parisod, C., Pfennig, K., Rice, A. M., Ritchie, M. G., Seifert, B., Smadja, C. M., Stelkens, R., Szymura, J. M., Väinölä, R., Wolf, J. B. W., and Zinner, D. 2013. Hybridization and speciation*. *Journal of Evolutionary Biology*, 26(2): 229–246.

Adams, A. M. and Hudson, R. R. 2004. Maximum-Likelihood Estimation of Demographic Parameters Using the Frequency Spectrum of Unlinked Single-Nucleotide Polymorphisms. *Genetics*, 168(3): 1699–1712.

Adavoudi, R. and Pilot, M. 2021. Consequences of Hybridization in Mammals: A Systematic Review. *Genes*, 13(1): 50.

Ali, O. A., O'Rourke, S. M., Amish, S. J., Meek, M. H., Luikart, G., Jeffres, C., and Miller, M. R. 2016. RAD capture (Rapture): Flexible and efficient sequence-based genotyping. *Genetics*, 202(2): 389–400.

Andermann, T., Fernandes, A. M., Olsson, U., Topel, M., Pfeil, B., Oxelman, B., Aleixo, A., Faircloth, B. C., and Antonelli, A. 2019. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Syst. Biol.*, 68(1): 32–46.

Andersen, L. N., Mailund, T., and Hobolth, A. 2014. Efficient computation in the IM model. *Journal of Mathematical Biology*, 68(6): 1423–1451.

Anderson, E. 1953. Introgressive Hybridization. *Biological Reviews*, 28(3): 280–307.

Anderson, E. and Hubricht, L. 1938. Hybridization in Tradescantia. III. The Evidence for Introgressive Hybridization. *American Journal of Botany*, 25(6): 396–402.

Andrade, P., Alves, J. M., Pereira, P., Rubin, C. J., Silva, E., Sprehn, C. G., Enbody, E., Afonso, S., Faria, R., Zhang, Y., Bonino, N., Duckworth, J. A., Garreau, H., Letnic, M., Strive, T., Thulin, C. G., Queney, G., Villafuerte, R., Jiggins, F. M., Ferrand, N., Andersson, L., and Carneiro, M. 2024. Selection against domestication alleles in introduced rabbit populations. *Nat. Ecol. Evol.*, 8(8): 1543–1555.

Arriola, P. E. and Ellstrand, N. C. 1996. Crop-To-Weed Gene Flow in the Genus Sorghum (Poaceae): Spontaneous Interspecific Hybridization between Johnsongrass, Sorghum halepense, and Crop Sorghum, S. Bicolor. *American Journal of Botany*, 83(9): 1153–1159.

Baack, E. J. and Rieseberg, L. H. 2007. A genomic view of introgression and hybrid speciation. *Current opinion in genetics & development*, 17(6): 513–518.

Baharian, S. and Gravel, S. 2018. On the decidability of population size histories from finite allele frequency spectra. *Theoretical Population Biology*, 120: 42–51.

Barton, N. H. 2006. Evolutionary biology: how did the human species form? *Curr. Biol.*, 16: R647–R650.

Becquet, C., Patterson, N., Stone, A. C., Przeworski, M., and Reich, D. 2007. Genetic Structure of Chimpanzee Populations. *PLOS Genetics*, 3(4): e66.

Beerli, P. and Felsenstein, J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98(8): 4563–4568.

Bhaskar, A. and Song, Y. S. 2014. DESCARTES' RULE OF SIGNS AND THE IDENTIFIABILITY OF POPULATION DEMOGRAPHIC MODELS FROM GENOMIC VARIATION DATA. *Annals of Statistics*, 42(6): 2469–2493.

Bi, K., Linderoth, T., Singhal, S., Vanderpool, D., Patton, J. L., Nielsen, R., Moritz, C., and Good, J. M. 2019. Temporal genomic contrasts reveal rapid evolutionary responses in an alpine mammal during recent climate change. *PLoS Genet.*, 15(5): e1008119.

Blischak, P. D., Chifman, J., Wolfe, A. D., and Kubatko, L. S. 2018. HyDe: a Python package for genome-scale hybridization detection. *Syst. Biol.*, 67(5): 821–829.

Bonnet, T., Leblois, R., Rousset, F., and Crochet, P.-A. 2017. A reassessment of explanations for discordant introgressions of mitochondrial and nuclear genomes. *Evolution*, 71(9): 2140–2158.

Brand, C. M., White, F. J., Rogers, A. R., and Webster, T. H. 2022. Estimating bonobo (Pan paniscus) and chimpanzee (Pan troglodytes) evolutionary history from nucleotide site patterns. *Proceedings of the National Academy of Sciences*, 119(17): e2200858119.

Brown, J. H. 1971. Mechanisms of competitive exclusion between two species of chipmunks. *Ecology*, 52(2): 305–311.

Brunsfeld, S. J., Soltis, D. E., and Soltis, P. S. 1992. Evolutionary Patterns and Processes in Salix Sect. Longifoliae: Evidence from Chloroplast DNA. *Systematic Botany*, 17(2): 239–256.

Burgess, R. and Yang, Z. 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. *Mol. Biol. Evol.*, 25: 1979–1994.

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Eichler, E. E., Korbel, J. O., Lee, C., Marschall, T., Devine, S. E., Harvey, W. T., Zhou, W., Mills, R. E., Rausch, T., Kumar, S., Alkan, C., Hormozdiari, F., Chong, Z., Chen, Y., Yang, X., Lin, J., Gerstein, M. B., Kai, Y., Zhu, Q., Yilmaz, F., Xiao, C., Paul Flicek, Germer, S., Brand, H., Hall, I. M., Talkowski, M. E., Narzisi, G., and Zody, M. C. 2022. High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, 185(18): 3426–3440.e19.

Cao, Z. and Nakhleh, L. 2019. Empirical Performance of Tree-based Inference of Phylogenetic Networks.

Caswell, J. L., Mallick, S., Richter, D. J., Neubauer, J., Schirmer, C., Gnerre, S., and Reich, D. 2008. Analysis of Chimpanzee History Based on Genome Sequence Alignments. *PLOS Genetics*, 4(4): e1000057.

Chan, Y.-C., Roos, C., Inoue-Murayama, M., Inoue, E., Shih, C.-C., and Vigilant, L. 2012. A comparative analysis of Y chromosome and mtDNA phylogenies of the Hylobates gibbons. *BMC Evolutionary Biology*, 12(1): 150.

Chifman, J. and Kubatko, L. 2014. Quartet inference from snp data under the coalescent model. *Bioinformatics*, 30(23): 3317–3324.

Chung, Y. and Hey, J. 2017. Bayesian Analysis of Evolutionary Divergence with Genomic Data under Diverse Demographic Models. *Molecular Biology and Evolution*, 34(6): 1517–1528.

Costa, R. J. and Wilkinson-Herbots, H. 2017. Inference of Gene Flow in the Process of Speciation: An Efficient Maximum-Likelihood Method for the Isolation-with-Initial-Migration Model. *Genetics*, 205(4): 1597–1618.

Dalquen, D., Zhu, T., and Yang, Z. 2017. Maximum likelihood implementation of an isolation-with-migration model for three species. *Syst. Biol.*, 66: 379–398.

Dalquest, W. W., Baskin, J., and Schultz, G. 1996. Fossil mammals from a late miocene (clarendonian) site in beaver county, oklahoma. *Contributions in Mammalogy: A Memorial Volume Honoring Dr. J. Knox Jones, Jr. Museum of Texas Tech University*, pages 107–137.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience*, 10(2): giab008.

de Manuel, M., Kuhlwilm, M., Frandsen, P., Sousa, V. C., Desai, T., Prado-Martinez, J., Hernandez-Rodriguez, J., Dupanloup, I., Lao, O., Hallast, P., Schmidt, J. M., Heredia-Genestar, J. M., Benazzo, A., Barbujani, G., Peter, B. M., Kuderna, L. F. K., Casals, F., Angedakin, S., Arandjelovic, M., Boesch, C., Kühl, H., Vigilant, L., Langergraber, K., Novembre, J., Gut, M., Gut, I., Navarro, A.,

Carlsen, F., Andrés, A. M., Siegismund, H. R., Scally, A., Excoffier, L., Tyler-Smith, C., Castellano, S., Xue, Y., Hvilsom, C., and Marques-Bonet, T. 2016. Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354(6311): 477–481.

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

Degnan, J. H. 2018. Modeling hybridization under the network multispecies coalescent. *Syst. Biol.*, 67(5): 786–799.

Dickey, J. M. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.*, 42(1): 204–223.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. 2011a. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, 28(8): 2239–2252.

Durand, E. Y., Patterson, N., Reich, D., and Slatkin, M. 2011b. Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.*, 28(8): 2239–2252.

Eaton, D. A. and Ree, R. H. 2013. Inferring phylogeny and introgression using RADseq data: an example from flowering plants (*Pedicularis: Orobanchaceae*). *Syst. Biol.*, 62(5): 689–706.

Edelman, N. B. and Mallet, J. 2021. Prevalence and Adaptive Impact of Introgression. *Annual Review of Genetics*, 55(Volume 55, 2021): 265–283.

Edelman, N. B., Frandsen, P. B., Miyagi, M., Clavijo, B., Davey, J., Dikow, R. B., Garcia-Accinelli, G., Van Belleghem, S. M., Patterson, N., Neafsey, D. E., Challis, R., Kumar, S., Moreira, G. R. P., Salazar, C., Chouteau, M., Counterman, B. A., Papa, R., Blaxter, M., Reed, R. D., Dasmahapatra, K. K., Kronforst, M., Joron, M., Jiggins, C. D., McMillan, W. O., Di Palma, F., Blumberg, A. J., Wakeley, J., Jaffe, D., and Mallet, J. 2019. Genomic architecture and introgression shape a butterfly radiation. *Science*, 366(6465): 594–599.

Edwards, S., Cloutier, A., and Baker, A. 2017. Conserved nonexonic elements: a novel class of marker for phylogenomics. *Syst. Biol.*, 66(6): 1028–1044.

Ellstrand, N. C. 2014. Is gene flow the most important evolutionary force in plants? *American Journal of Botany*, 101(5): 737–753.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., and Foll, M. 2013. Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics*, 9(10): e1003905.

Excoffier, L., Marchi, N., Marques, D. A., Matthey-Doret, R., Gouy, A., and Sousa, V. C. 2021. fastsimcoal2: demographic inference under complex evolutionary scenarios. *Bioinformatics*, 37(24): 4882–4885.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., and Glenn, T. C. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.*, 61(5): 717–726.

Feder, J. L., Egan, S. P., and Nosil, P. 2012. The genomics of speciation-with-gene-flow. *Trends in genetics: TIG*, 28(7): 342–350.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17(6): 368–376.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10): 2585–2593.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2020. A Bayesian implementation of the multispecies coalescent model with introgression for phylogenomic analysis. *Mol. Biol. Evol.*, 37(4): 1211–1223.

Flouri, T., Huang, J., Jiao, X., Kapli, P., Rannala, B., and Yang, Z. 2022. Bayesian phylogenetic inference using relaxed-clocks and the multispecies coalescent. *Mol. Biol. Evol.*, 39(8): msac161.

Flouri, T., Jiao, X., Huang, J., Rannala, B., and Yang, Z. 2023. Efficient Bayesian inference under the multispecies coalescent with migration. *Proc. Nat. Acad. Sci. U.S.A.*, 120(44): e2310708120.

Folk, R. A., Soltis, P. S., Soltis, D. E., and Guralnick, R. 2018. New prospects in the detection and comparative analysis of hybridization in the tree of life. *American Journal of Botany*, 105(3): 364–375.

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).

Forsythe, E. S., Sloan, D. B., and Beilstein, M. A. 2020. Divergence-Based Introgression Polarization. *Genome Biology and Evolution*, 12(4): 463–478.

Fumagalli, M. 2013. Assessing the effect of sequencing depth and sample size in population genetics inferences. *PLoS One*, 8: e79667.

Gelman, A. and Meng, X. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13: 163–185.

Good, J. M. and Sullivan, J. 2001. Phylogeography of the red-tailed chipmunk (*Tamias ruficaudus*), a northern Rocky Mountain endemic. *Mol. Ecol.*, 10(11): 2683–2695.

Good, J. M., Demboski, J. R., Nagorsen, D. W., and Sullivan, J. 2003. Phylogeography and introgressive hybridization: chipmunks (genus *Tamias*) in the northern Rocky Mountains. *Evolution*, 57(8): 1900–1916.

Good, J. M., Hird, S., Reid, N., Demboski, J. R., Steppan, S. J., Martin-Nims, T. R., and Sullivan, J. 2008. Ancient hybridization and mitochondrial capture between two species of chipmunks. *Mol. Ecol.*, 17(5): 1313–1327.

Green, P. 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82: 711–732.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., Hansen, N. F., Durand, E. Y., Malaspinas, A.-S., Jensen, J. D., Marques-

Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H. A., Good, J. M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E. S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V. B., Golovanova, L. V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R. W., Johnson, P. L. F., Eichler, E. E., Falush, D., Birney, E., Mullikin, J. C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D., and Pääbo, S. 2010. A Draft Sequence of the Neandertal Genome. *Science*, 328(5979): 710–722.

Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G., and Siepel, A. 2011. Bayesian inference of ancient human demography from individual genome sequences. *Nature Genet.*, 43: 1031–1034.

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., and Bustamante, C. D. 2009. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS Genetics*, 5(10): e1000695.

Hallast, P., Maisano Delser, P., Batini, C., Zadik, D., Rocchi, M., Schempp, W., Tyler-Smith, C., and Jobling, M. A. 2016. Great ape Y Chromosome and mitochondrial DNA phylogenies reflect subspecies structure and patterns of mating and dispersal. *Genome Research*, 26(4): 427–439.

Heiser, C. B. 1947. Hybridization Between the Sunflower Species Helianthus annuus and H. petiolaris. *Evolution*, 1(4): 249–262.

Heled, J. and Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, 27: 570–580.

Heller, H. C. 1971. Altitudinal zonation of chipmunks (*Eutamias*): interspecific aggression. *Ecology*, 52(2): 312–319.

Hey, J. 2010a. The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. *Molecular Biology and Evolution*, 27(4): 921–933.

Hey, J. 2010b. Isolation with migration models for more than two populations. *Mol. Biol. Evol.*, 27(4): 905–920.

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.

Hey, J. and Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8): 2785–2790.

Hey, J., Chung, Y., Sethuraman, A., Lachance, J., Tishkoff, S., Sousa, V. C., and Wang, Y. 2018. Phylogeny estimation by integration over isolation with migration models. *Mol. Biol. Evol.*, 35(11): 2805—-2818.

Hibbins, M. S. and Hahn, M. W. 2019. The Timing and Direction of Introgression Under the Multispecies Network Coalescent. *Genetics*, 211(3): 1059–1073.

Hibbins, M. S. and Hahn, M. W. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*, 220(2): iyab173.

Hird, S., Reid, N., Demboski, J., and Sullivan, J. 2010. Introgression at differentially aged hybrid zones in red-tailed chipmunks. *Genetica*, 138(8): 869–883.

Hobolth, A., Andersen, L. N., and Mailund, T. 2011. On Computing the Coalescence Time Density in an Isolation-With-Migration Model With Few Samples. *Genetics*, 187(4): 1241–1243.

Huang, J., Flouri, T., and Yang, Z. 2020. A simulation study to examine the information content in phylogenomic datasets under the multispecies coalescent model. *Mol. Biol. Evol.*, 37(11): 3211–3224.

Huang, J., Thawornwattana, Y., Flouri, T., Mallet, J., and Yang, Z. 2022a. Inference of gene flow between species under misspecified models. *Mol. Biol. Evol.*, 39(12): msac237.

Huang, J., Bennett, J., Flouri, T., and Yang, Z. 2022b. Phase resolution of heterozygous sites in diploid genomes is important to phylogenomic analysis under the multispecies coalescent model. *Syst. Biol.*, 71(2): 334–352.

Inoue, E., Inoue-Murayama, M., Vigilant, L., Takenaka, O., and Nishida, T. 2008. Relatedness in wild chimpanzees: Influence of paternity, male philopatry, and demographic factors. *American Journal of Physical Anthropology*, 137(3): 256–262.

Jackson, N. D., Morales, A. E., Carstens, B. C., and O'Meara, B. C. 2017. PHRAPL: Phylogeographic Inference Using Approximate Likelihoods. *Systematic Biology*, 66(6): 1045–1053.

Jeffreys, H. 1939. *Theory of Probability*. Clarendon Press, Oxford, England.

Ji, J., Jackson, D. J., Leache, A. D., and Yang, Z. 2023. Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipmunks. *Syst. Biol.*, 72(2): 446–465.

Jiao, X., Flouri, T., and Yang, Z. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.*, page 10.1093/nsr/nwab127.

Jones, G. R. 2019. Divergence Estimation in the Presence of Incomplete Lineage Sorting and Migration. *Systematic Biology*, 68(1): 19–31.

Jorde, L. B. 2000. Linkage Disequilibrium and the Search for Complex Disease Genes. *Genome Research*, 10(10): 1435–1444.

Jouganous, J., Long, W., Ragsdale, A. P., and Gravel, S. 2017. Inferring the Joint Demographic History of Multiple Populations: Beyond the Diffusion Approximation. *Genetics*, 206(3): 1549–1567.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–123. Academic Press, New York.

Karin, B. R., Gamble, T., and Jackman, T. R. 2020. Optimizing phylogenomics with rapidly evolving long exons: Comparison with anchored hybrid enrichment and ultraconserved elements. *Mol. Biol. Evol.*, 37(3): 904–922.

Kim, B. Y., Gellert, H. R., Church, S. H., Suvorov, A., Anderson, S. S., Barmina, O., Beskid, S. G., Comeault, A. A., Crown, K. N., Diamond, S. E., Dorus, S., Fujichika, T., Hemker, J. A., Hrcek,

J., Kankare, M., Katoh, T., Magnacca, K. N., Martin, R. A., Matsunaga, T., Medeiros, M. J., Miller, D. E., Pitnick, S., Simoni, S., Steenwinkel, T. E., Schiffer, M., Syed, Z. A., Takahashi, A., Wei, K. H., Yokoyama, T., Eisen, M. B., Kopp, A., Matute, D., Obbard, D. J., O'Grady, P. M., Price, D. K., Toda, M. J., Werner, T., and Petrov, D. A. 2023. Single-fly assemblies fill major phylogenomic gaps across the *Drosophilidae* Tree of Life. *bioRxiv*.

Kimura, M. 1964. Diffusion models in population genetics. *Journal of Applied Probability*, 1(2): 177–232.

Kingman, J. F. C. 1982. The coalescent. *Stochastic Processes and their Applications*, 13(3): 235–248.

Korneliussen, T. S., Albrechtsen, A., and Nielsen, R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1): 356.

Kubatko, L. 2019. The multispecies coalescent. In D. Balding, I. Moltke, and J. Marioni, editors, *Handbook of Statistical Genomics*, pages 219–245. Wiley, New York, 4th edition.

Kubatko, L. S. and Chifman, J. 2019. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evol. Biol.*, 19(1): 112.

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Marques-Bonet, T., Andrés, A. M., Viola, B., Pääbo, S., Meyer, M., Siepel, A., and Castellano, S. 2016a. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, 530(7591): 429–433.

Kuhlwilm, M., de Manuel, M., Nater, A., Greminger, M. P., Krützen, M., and Marques-Bonet, T. 2016b. Evolution and demography of the great apes. *Current Opinion in Genetics & Development*, 41: 124–129.

Kuhlwilm, M., Han, S., Sousa, V. C., Excoffier, L., and Marques-Bonet, T. 2019. Ancient admixture from an extinct ape lineage into bonobos. *Nature Ecology & Evolution*, 3(6): 957–965.

Lartillot, N. and Philippe, H. 2006. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55: 195–207.

Leaché, A. D., Harris, R. B., Rannala, B., and Yang, Z. 2014. The influence of gene flow on species tree estimation: a simulation study. *Syst. Biol.*, 63(1): 17–30.

Lemmon, A. R., Emme, S. A., and Lemmon, E. M. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst. Biol.*, 61(5): 727–744.

Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21): 2987–2993.

Li, H. and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25: 1754–1760.

Li, L., Comi, T. J., Bierman, R. F., and Akey, J. M. 2024. Recurrent gene flow between Neanderthals and modern humans over the past 200,000 years. *Science*, 385(6705): eadi1768.

Lindley, D. 1957. A statistical paradox. *Biometrika*, 44: 187–192.

Liu, K. J., Dai, J., Truong, K., Song, Y., Kohn, M. H., and Nakhleh, L. 2014. An HMM-Based Comparative Genomic Framework for Detecting Introgression in Eukaryotes. *PLOS Computational Biology*, 10(6): e1003649. Publisher: Public Library of Science.

Lobon, I., Tucci, S., de Manuel, M., Ghirotto, S., Benazzo, A., Prado-Martinez, J., Lorente-Galdos, B., Nam, K., Dabad, M., Hernandez-Rodriguez, J., Comas, D., Navarro, A., Schierup, M. H., Andres, A. M., Barbujani, G., Hvilsom, C., and Marques-Bonet, T. 2016. Demographic History of the Genus Pan Inferred from Whole Mitochondrial Genome Reconstructions. *Genome Biology and Evolution*, 8(6): 2020–2030.

Lohse, K. and Frantz, L. A. 2014. Neandertal admixture in Eurasia confirmed by maximum-likelihood analysis of three genomes. *Genetics*, 196(4): 1241–1251.

Lou, D. I., Hussmann, J. A., McBee, R. M., Acevedo, A., Andino, R., Press, W. H., and Sawyer, S. L. 2013. High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, 110(49): 19872–19877.

Maier, R., Flegontov, P., Flegontova, O., Işıldak, U., Changmai, P., and Reich, D. 2023. On the limits of fitting complex models of population history to f-statistics. *eLife*, 12: e85492.

Malinsky, M., Challis, R. J., Tyers, A. M., Schiffels, S., Terai, Y., Ngatunga, B. P., Miska, E. A., Durbin, R., Genner, M. J., and Turner, G. F. 2015. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, 350(6267): 1493–1498.

Malinsky, M., Svardal, H., Tyers, A. M., Miska, E. A., Genner, M. J., Turner, G. F., and Durbin, R. 2018. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nat. Ecol. Evol.*, 2(12): 1940–1955.

Malinsky, M., Matschiner, M., and Svardal, H. 2021. Dsuite - Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, 21(2): 584–595.

Mallet, J. 2005. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5): 229–237.

Mallet, J. 2007. Hybrid speciation. *Nature*, 446(7133): 279–283.

Mallet, J., Besansky, N., and Hahn, M. W. 2016. How reticulated are species? *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, 38(2): 140–149.

Marques, D. A., Meier, J. I., and Seehausen, O. 2019. A Combinatorial View on Speciation and Adaptive Radiation. *Trends in Ecology & Evolution*, 34(6): 531–544.

Martin, S. H. and Jiggins, C. D. 2017. Interpreting the genomic landscape of introgression. *Current Opinion in Genetics & Development*, 47: 69–74.

Martin, S. H., Davey, J. W., and Jiggins, C. D. 2015. Evaluating the Use of ABBA–BABA Statistics to Locate Introgressed Loci. *Molecular Biology and Evolution*, 32(1): 244–257.

Mayr, E. 1942. *Systematics and the Origin of Species*. Columbia University Press.

Mayr, E. 1963. *Animal Species and Evolution*. Harvard University Press.

Mayr, E. 1970. *Populations, Species, and Evolution: An Abridgment of Animal Species and Evolution*. Harvard University Press.

McBrearty, S. and Jablonski, N. G. 2005. First fossil chimpanzee. *Nature*, 437(7055): 105–108.

McElroy, K., Black, A., Dolman, G., Horton, P., Pedler, L., Campbell, C. D., Drew, A., and Joseph, L. 2020. Robbery in progress: Historical museum collections bring to light a mitochondrial capture within a bird species widespread across southern Australia, the copperback quail-thrush *Cinclosoma clarum*. *Ecol. Evol.*, 10(13): 6785–6793.

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

Mo, Z. and Siepel, A. 2023. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLOS Genetics*, 19(11): e1011032. Publisher: Public Library of Science.

Molloy, E. K. and Warnow, T. 2019. TreeMerge: a new method for improving the scalability of species tree estimation methods. *Bioinformatics*, 35(14): i417–i426.

Molloy, E. K., Durvasula, A., and Sankararaman, S. 2021. Advancing admixture graph estimation via maximum likelihood network orientation. *Bioinformatics (Oxford, England)*, 37(Suppl_1): i142–i150.

Myers, S., Fefferman, C., and Patterson, N. 2008. Can one learn history from the allelic spectrum? *Theoretical Population Biology*, 73(3): 342–348.

Nath, H. B. and Griffiths, R. C. 1993. The coalescent in two colonies with symmetric migration. *Journal of Mathematical Biology*, 31(8): 841–851.

Nielsen, R. 2000. Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms. *Genetics*, 154(2): 931–942.

Nielsen, R. and Wakeley, J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, 158: 885–896.

Noor, M. A., Johnson, N. A., and Hey, J. 2000. Gene flow between *Drosophila pseudoobscura* and *D. persimilis*. *Evolution*, 54: 2174–2175; discussion 2176–2177.

Notohara, M. 1990. The coalescent and the genealogical process in geographically structured population. *Journal of Mathematical Biology*, 29(1): 59–75.

Ogilvie, H. A., Bouckaert, R. R., and Drummond, A. J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Mol. Biol. Evol.*, 34(8): 2101–2114.

O'Hagan, A. and Forster, J. 2004. *Kendall's Advanced Theory of Statistics: Bayesian Inference*. Arnold, London.

Osada, N. and Wu, C.-I. 2005. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics*, 169(1): 259–264.

Pang, X. X. and Zhang, D. Y. 2024. Detection of ghost introgression requires exploiting topological and branch length information. *Syst. Biol.*, 73(1): 207–222.

Patterson, B. D. and Thaeler Jr, C. S. 1982. The mammalian baculum: hypotheses on the nature of bacular variability. *J. Mammal.*, 63(1): 1–15.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., and Reich, D. 2012. Ancient Admixture in Human History. *Genetics*, 192(3): 1065–1093.

Pease, J. B. and Hahn, M. W. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, 64(4): 651–662.

Pickrell, J. K. and Pritchard, J. K. 2012. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS Genetics*, 8(11): e1002967.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Research*, 20(1): 110–121.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., *et al.* 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*, page 201178.

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., Veeramah, K. R., Woerner, A. E., O'Connor, T. D., Santpere, G., Cagan, A., Theunert, C., Casals, F., Laayouni, H., Munch, K., Hobolth, A., Halager, A. E., Malig, M., Hernandez-Rodriguez, J., Hernando-Herraez, I., Prüfer, K., Pybus, M., Johnstone, L., Lachmann, M., Alkan, C., Twigg, D., Petit, N., Baker, C., Hormozdiari, F., Fernandez-Callejo, M., Dabad, M., Wilson, M. L., Stevison, L., Camprubí, C., Carvalho, T., Ruiz-Herrera, A., Vives, L., Mele, M., Abello, T., Kondova, I., Bontrop, R. E., Pusey, A., Lankester, F., Kiyang, J. A., Bergl, R. A., Lonsdorf, E., Myers, S., Ventura, M., Gagneux, P., Comas, D., Siegismund, H., Blanc, J., Agueda-Calpena, L., Gut, M., Fulton, L., Tishkoff, S. A., Mullikin, J. C., Wilson, R. K., Gut, I. G., Gonder, M. K., Ryder, O. A., Hahn, B. H., Navarro, A., Akey, J. M., Bertranpetit, J., Reich, D., Mailund, T., Schierup, M. H., Hvilsom, C., Andrés, A. M., Wall, J. D., Bustamante, C. D., Hammer, M. F., Eichler, E. E., and Marques-Bonet, T. 2013. Great ape genetic diversity and population history. *Nature*, 499(7459): 471–475.

Ragsdale, A. P. and Gravel, S. 2019. Models of archaic admixture and recent history from two-locus statistics. *PLOS Genetics*, 15(6): e1008204.

Ragsdale, A. P. and Gravel, S. 2020. Unbiased Estimation of Linkage Disequilibrium from Unphased Data. *Molecular Biology and Evolution*, 37(3): 923–932.

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

Rannala, B. and Yang, Z. 2017. Efficient bayesian species tree inference under the multispecies coalescent. *Syst. Biol.*, 66: 823–842.

Rannala, B., Edwards, S. V., Leaché, A. D., and Yang, Z. 2020. The multispecies coalescent model and species tree inference. In C. Scornavacca, F. Delsuc, and N. Galtier, editors, *Phylogenetics in the Genomic Era*, chapter 3.3, page 3.3:1–20. No Commercial Publisher.

Reid, N., Demboski, J. R., and Sullivan, J. 2012. Phylogeny estimation of the radiation of western north american chipmunks (*Tamias*) in the face of introgression using reproductive protein genes. *Syst. Biol.*, 61(1): 44.

Ribas, G., Milne, R. L., Gonzalez-Neira, A., and Benítez, J. 2008. Haplotype patterns in cancer-related genes with long-range linkage disequilibrium: no evidence of association with breast cancer or positive selection. *European Journal of Human Genetics*, 16(2): 252–260.

Rieseberg, L. H. 1997. Hybrid Origins of Plant Species. *Annual Review of Ecology, Evolution, and Systematics*, 28(Volume 28, 1997): 359–389.

Rogers, A. R. 2019. Legofit: estimating population history from genetic data. *BMC Bioinformatics*, 20(1): 526.

Root, J. J., Calisher, C. H., and Beaty, B. J. 2001. Microhabitat partitioning by two chipmunk species (*Tamias*) in western Colorado. *Western North American Naturalist*, pages 114–118.

Rubin, B. E., Ree, R. H., and Moreau, C. S. 2012. Inferring phylogenies from rad sequence data. *PLoS One*, 7(4): e33394.

Sarver, B. A., Demboski, J. R., Good, J. M., Forshee, N., Hunter, S. S., and Sullivan, J. 2017. Comparative phylogenomic assessment of mitochondrial introgression among several species of chipmunks (*Tamias*). *Genome Biol. Evol.*, 9(1): 7–19.

Sarver, B. A. J., Herrera, N. D., Sneddon, D., Hunter, S. S., Settles, M. L., Kronenberg, Z., Demboski, J. R., Good, J. M., and Sullivan, J. 2021. Diversification, introgression, and rampant cytonuclear discordance in Rocky Mountains chipmunks (Sciuridae: *Tamias*). *Syst. Biol.*, 70(5): 908–921.

Schrider, D. R. and Kern, A. D. 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, 34(4): 301–312.

Self, S. and Liang, K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82: 605–610.

Sexton, J. P., Clemens, M., Bell, N., Hall, J., Fyfe, V., and Hoffmann, A. A. 2024. Patterns and effects of gene flow on adaptation across spatial scales: implications for management. *Journal of Evolutionary Biology*, 37(6): 732–745.

Shi, C. and Yang, Z. 2018. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, 35: 159–179.

Silverman, B. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Slatkin, M. 1985. Gene Flow in Natural Populations. *Annual Review of Ecology, Evolution, and Systematics*, 16(Volume 16,): 393–430.

Slatkin, M. 1987. Gene Flow and the Geographic Structure of Natural Populations. *Science*, 236(4803): 787–792.

Smith, M. L. and Hahn, M. W. 2024. Selection leads to false inferences of introgression using popular methods. *Genetics*, 227(4): iyae089.

Solis-Lemus, C. and Ane, C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.*, 12(3): e1005896.

Solis-Lemus, C., Bastide, P., and Ane, C. 2017. PhyloNetworks: a package for phylogenetic networks. *Mol. Biol. Evol.*, 34(12): 3292–3298.

Soltis, P. S. and Soltis, D. E. 2009. The role of hybridization in plant speciation. *Annual Review of Plant Biology*, 60: 561–588.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30: 1312–1313.

Steinrücken, M., Spence, J. P., Kamm, J. A., Wieczorek, E., and Song, Y. S. 2018. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. *Molecular Ecology*, 27(19): 3873–3888. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.14565.

Stephens, M. and Donnelly, P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *American Journal of Human Genetics*, 73(5): 1162–1169.

Stephens, M., Smith, N. J., and Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. *American Journal of Human Genetics*, 68(4): 978–989.

Sullivan, J., Demboski, J., Bell, K., Hird, S., Sarver, B., Reid, N., and Good, J. 2014. Divergence with gene flow within the recent chipmunk radiation (*Tamias*). *Heredity*, 113(3): 185–194.

Suvorov, A., Kim, B. Y., Wang, J., Armstrong, E. E., Peede, D., D'Agostino, E. R. R., Price, D. K., Waddell, P., Lang, M., Courtier-Orgogozo, V., David, J. R., Petrov, D., Matute, D. R., Schrider, D. R., and Comeault, A. A. 2022. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Curr. Biol.*, 32: 111–123.

Swofford, D. L. 2003. *PAUP*: Phylogenetic Analysis by Parsimony (*and Other Methods), Version 4*. Sinauer Associates, Sanderland, Massachusetts.

Takahata, N. 1988. The coalescent in two partially isolated diffusion populations. *Genetical Research*, 52(3): 213–222.

Takahata, N., Satta, Y., and Klein, J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.*, 48: 198–221.

Tavaré, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect. Math. Life Sci.*, 17: 57–86.

Taylor, S. A. and Larson, E. L. 2019. Insights from genomes into the evolutionary importance and prevalence of hybridization in nature. *Nature Ecology & Evolution*, 3(2): 170–177.

Terhorst, J. and Song, Y. S. 2015. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. *Proceedings of the National Academy of Sciences of the United States of America*, 112(25): 7677–7682.

Thawornwattana, Y., Dalquen, D., and Yang, Z. 2018. Coalescent analysis of phylogenomic data confidently resolves the species relationships in the *Anopheles gambiae* species complex. *Mol. Biol. Evol.*, 35(10): 2512–2527.

Thawornwattana, Y., Seixas, F. A., Mallet, J., and Yang, Z. 2022. Full-likelihood genomic analysis clarifies a complex history of species divergence and introgression: the example of the erato-sara group of Heliconius butterflies. *Syst. Biol.*, 71(5): 1159–1177.

Thawornwattana, Y., Huang, J., Flouris, T., Mallet, J., and Yang, Z. 2023a. Inferring the direction of introgression using genomic sequence data. *Mol. Biol. Evol.*, 40(8): msad178.

Thawornwattana, Y., Seixas, F., Yang, Z., and Mallet, J. 2023b. Major patterns in the introgression history of Heliconius butterflies. *ELife*, 12: RP90656.

Thawornwattana, Y., Flouris, T., Mallet, J., and Yang, Z. 2024. Inference of continuous gene flow between species under misspecified models. *Mol. Biol. Evol.*, (submitted).

Tiley, G. P., Flouri, T., Jiao, X., Poelstra, J. P., Xu, B., Zhu, T., Rannala, B., Yoder, A. D., and Yang, Z. 2023. Estimation of species divergence times in presence of cross-species gene flow. *Syst. Biol.*, 72(4): 820–836.

Verdinelli, I. and Wasserman, L. 1995. Computing bayes factors using a generalization of the savage-dickey density ratio. *J. Am. Stat. Assoc.*, 90(430): 614–618.

Wang, Y. and Hey, J. 2010. Estimating divergence parameters with small samples from a large number of loci. *Genetics*, 184: 363–379.

Wegmann, D. and Excoffier, L. 2010. Bayesian Inference of the Demographic History of Chimpanzees. *Molecular Biology and Evolution*, 27(6): 1425–1435.

Wen, D. and Nakhleh, L. 2018. Coestimating Reticulate Phylogenies and Gene Trees from Multilocus Sequence Data. *Systematic Biology*, 67(3): 439–457.

Wen, D., Yu, Y., and Nakhleh, L. 2016. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLOS Genetics*, 12(5): e1006006.

Wen, D., Yu, Y., Zhu, J., and Nakhleh, L. 2018. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.*, 67(4): 735–740.

White, J. A. 2010. *The Baculum in the Chipmunks of Western North America*. Good Press.

Wilkinson-Herbots, H. M. 1998. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6): 535–585.

Won, Y.-J. and Hey, J. 2005. Divergence population genetics of chimpanzees. *Molecular Biology and Evolution*, 22(2): 297–307.

Wright, S. 1931. EVOLUTION IN MENDELIAN POPULATIONS. *Genetics*, 16(2): 97–159.

Xu, B. and Yang, Z. 2016. Challenges in species tree estimation under the multispecies coalescent model. *Genetics*, 204(4): 1353–1368.

Yang, Y., Pang, X.-X., Ding, Y.-M., Zhang, B.-W., Bai, W.-N., and Zhang, D.-Y. 2025. Synergizing Bayesian and Heuristic Approaches: D-BPP Uncovers Ghost Introgression in Panthera and Thuja.

Yang, Z. 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39: 105–111.

Yang, Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, 42: 294–307.

Yang, Z. 1998. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.*, 47: 125–133.

Yang, Z. 2006. *Computational Molecular Evolution*. Oxford University Press, Oxford, UK.

Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5): 854–865.

Yang, Z. and Flouri, T. 2022. Estimation of cross-species introgression rates using genomic data despite model unidentifiability. *Mol. Biol. Evol.*, 39(5): 10.1093/molbev/msac083.

Yang, Z. and Rannala, B. 2010. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. U.S.A.*, 107: 9264–9269.

Yang, Z. and Rannala, B. 2014. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.*, 31(12): 3125–3135.

Yu, Y. and Nakhleh, L. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10): S10.

Yu, Y., Degnan, J. H., and Nakhleh, L. 2012. The Probability of a Gene Tree Topology within a Phylogenetic Network with Applications to Hybridization Detection. *PLOS Genetics*, 8(4): e1002660.

Yu, Y., Ristic, N., and Nakhleh, L. 2013. Fast algorithms and heuristics for phylogenomics under ILS and hybridization. *BMC Bioinformatics*, 14(15): S6.

Yu, Y., Dong, J., Liu, K. J., and Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. *Proceedings of the National Academy of Sciences*, 111(46): 16448–16453.

Zhang, C. and Nielsen, R. 2025. Waster: Practical *de novo* phylogenomics from low-coverage short reads. *BioRxiv*, page 10.1101/2025.01.20.633983.

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.

Zhang, C., Nielsen, R., and Mirarab, S. 2025. Caster: Direct species tree inference from whole-genome alignments. *Science*, page eadk9688.

Zheng, Y. and Janke, A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics*, 19(1): 10.

Zhu, T. and Yang, Z. 2012. Maximum likelihood implementation of an isolation-with-migration model with three species for testing speciation with gene flow. *Molecular Biology and Evolution*, 29(10): 3131–3142.

Zhu, T. and Yang, Z. 2021. Complexity of the simplest species tree problem. *Mol. Biol. Evol.*, 39: 3993–4009.

Zhu, T., Flouri, T., and Yang, Z. 2022. A simulation study to examine the impact of recombination on phylogenomic inferences under the multispecies coalescent model. *Mol. Ecol.*, 31: 2814–2829.