Geographies and Genealogies: Phylogeographic Simulation and Bayesian Approaches to Statistical Phylogeographic Model Selection

BY

Jeet Sukumaran

Submitted to the graduate degree program in Ecology and Evolutionary Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Chairperson Dr. Mark T. Holder

Chairperson Dr. Rafe Brown

Dr. John Kelly

Dr. A. Townsend Peterson

Dr. Jorge Soberón

Dr. Xingong Li

Date Defended: Jan 31, 2012

The Disssertation Committee for Jeet Sukumaran certifies that this is the approved version of the following dissertation:

Geographies and Genealogies: Phylogeographic Simulation and Bayesian Approaches to Statistical Phylogeographic Model Selection

Chairperson Dr. Mark T. Holder

Chairperson Dr. Rafe Brown

Date Approved: Jan 31, 2012

Part I

Abstract

Abstract

A wide class of biogeographic or phylogeographic studies predicts the simultaneous divergence of co-distributed taxa. Typically, a geological event, or a climate-related change in geography, is hypothesized to have structured a broad range of biota, many components of which may only be distantly related to each other. Direct assessment of these predictions is precluded in many studies by the lack or paucity of appropriate fossils for calibration when estimating divergence times in a phylogenetic context. However, even without direct divergence time estimation of all the relevant splits, there might be sufficient information in the data to estimate the probability that these groups diverged simultaneously if the datasets are treated in a parallel, coordinated, and integrated fashion, rather than independently. This study investigates the statistical framework and methods used to address this issue.

Most current statistical phylogeographic methods rely on the coalescent as an underlying model. While the coalescent is robust to a range of violations of some of its assumptions, such as the Wright-Fisher demographic model, and, morever, has been elaborated or extended to allow the relaxing of some of its other assumptions, little has been done to assess and quanitfy how violations of these assumptions affect phylogeographic analysis in general, and phylogeographic model selection in particular. One of the major problems in evaluating the performance of phylogeographic methods with respect to their responses or behavior when the assumptions of the coalescent are violated is the lack of a rich or flexible non-coalesccent based spatially-explicit simulation engine. The first chapter of my dissertation is thus focussed on developing and producing such a simulator: a forward-time, agent-based, spatially-explicit simulation program that generates genealogies for multiple loci evolving in populations of multiple sexual diploid species on a spatio-temporally environmentally-heterogenous landscape.

The second chapter of the dissertation assesses the performance of an Approximate Bayesian Computation approach to simultaneous divergence time testing model selection. It profiles the performance this approach under a variety of conditions, ranging from ones in which its model assumptions are completely met, to ones in which they are selectively violated in varying degrees. While there currently are no full- or exact-likelihood methods that address this question, under the special controlled circumstances of the study it was possible to adapt an existing program to provide some indication of how a full-likehood method may work in contrast.

The third chapter of this work presents a program that simultaneously estimates the divergence time between sister populations of multiple species in parallel. This program uses a Bayesian statistical framework to analyze data from multiple genetic loci, integrating over uncertainty in gene trees, divergence times, and demographic parameters. If limited to two species, the program allows for reverse-jump MCMC to sample from models of different dimensionality with respect to the divergence time, so as to explicitly estimate the posterior probability of simultaneous divergence vs. non-simultaneous divergence.

Part II

Acknowledgements

Acknowledgements

This work would not be possible without the love and support of my parents, whose patience and understanding in some very trying and difficult times have and always will inspire and comfort me.

Both Mark T. Holder and Rafe Brown were much more than advisors and colleagues. They were, and shall always remain, first and foremost, friends. The best of the work presented here is the direct result of their guidance, input, advice and intervention.

The camaraderie, openess and scientific passion of both my fellow graduate students as well as the faculty were critical in fostering an environment where not only learning, but also personal and intellectual growth, were encouraged and cultivated.

The very many dogs and cats with whom I interacted over the past several were also a great source of joy, and helped me keep my sanity when things got difficult.

Part III

Table of Contents

Contents

Part IV

Dissertation

Chapter 1

Ginkgo: Spatially-Explicit Simulator of Complex Phylogeographic Histories

1.1 Introduction

While phylogeography has been an active discipline of evolutionary biology since the 1990's (*cf.* Avise, 2000), the field has seen dramatic changes recently. Knowles and Maddison (2002) made the case for the use of rigorous statistical approaches to phylogeographic studies, and the number of statistical methods and software for the analysis of interactions between the geographical and demographical history of populations and the corresponding genealogies has steadily grown (e.g., Carstens et al., 2004, 2005; Hickerson et al., 2007; Nielsen and Beaumont, 2009). The relative merits of these approaches have been discussed and debated in the literature (e.g., Garrick et al., 2008; Knowles, 2008; Templeton, 2009, 2008), from analytical (e.g., Beaumont et al., 2010) and simulation perspectives (e.g., Panchal, 2007; Panchal et al., 2007). In most cases where methods have been tested using simulated data, the simulation models have been based on the coalescent (Kingman, 1982a), or have been simple relative to real-world processes (e.g., Petit, 2008; Panchal et al., 2007). Simple simulations are easy to interpret and allow us to compare methods in the case of clean data, but there is a danger that the results of such studies may not be applicable to the analysis of real data.

Here we present `Ginkgo`, a C++ program for the agent-based simulation of genealogies of mul-

tiple independent diploid and haploid loci evolving in populations of multiple species in a spatially-explicit framework with dynamic geographies and environmental selection regimes. These sampled genealogies can be used directly with phylogeographical analyses that take phylogenetic trees as input, or sequence data can be simulated on the genealogies to produce input for phylogeographical software that operates on sequence data. `Ginkgo`'s simulation engine provides for complex multi-scale geographical sub-structuring as well as important population genetic and evolutionary processes (such as selection) that, while not typically accommodated during phylogeographic inference, are nonetheless commonly encountered in real-world data. `Ginkgo` thus allows for the design of more realistic and challenging tests of the performance of phylogeographic analysis methods, and to evaluate the robustness of different inference procedures to violation of their simplifying assumptions.

1.2 The landscape

The spatio-environmental framework of `Ginkgo` is represented by the "landscape", an abstract $n \times m$ rectangular grid of cells (with n and m determined by the user). Each cell is associated with a vector of environmental parameters or factors. The fitness of an organism in any particular cell is given by a function of these environmental factors and the organism's phenotype (see below). This fitness determines the organism's probability of survival in that cell.

Each cell has a carrying capacity associated which determines the maximum number of organisms from across all species that it can support. If the total number of organisms in a cell exceeds the cell's carrying capacity, then the organisms are ranked in order of their fitness in that cell (see below), and the lowest-ranked organisms are culled until the cell is at its carrying capacity.

Sexual reproduction is panmictic *within* a cell. Spatial-structuring arises through the control of movement of organisms between cells. Organisms move randomly between adjacent cells, but this migration is modulated by an entry cost associated with moving into a particular cell. The entry costs differ across cells and between species, to reflect different ecological constraints or vagilities. High entry costs can also be used to mimic barrier to gene flow between different regions of the landscape.

The migration phase begins with each organism being assigned a "movement capacity". This

movement capacity is determined by the user for each particular species, and can be specified as a fixed value, a value drawn from a parametric distribution (e.g., Poisson, normal or uniform), or a custom vector of probability values associated with different movement capacities. The organism then selects one of the nine cells that constitute its immediate neighborhood (i.e., its current cell as well as the the eight cells bordering it) with uniform random probability. The organism then attempts to "pay" the entry cost for that cell by deducting the cost from its movement capacity. If the organism has a non-negative movement capacity remaining, then it successfully moves into that destination cell. This process then repeats with the new cell taken as the current cell, until the organism's movement currency is depleted to 0 or less and it can no longer move into any cell, at which point the migration phase for that organism terminates.

Long-distance dispersal can be introduced at any point during the simulation, determined in advance by the user specifying a source cell, a destination cell, and the probability of dispersal.

A cell's entry costs, environmental factors, long-distance dispersal probabilities, and carrying capacity can be changed during the course of the simulation by specifying a schedule for these parameters. This allows one to model changes in climate or geological connectivity of the landscape.

1.3 Species

Organisms in the simulation are organized into classes of distinct ecologies, i.e., "species" or "lineages". Membership in a particular species determines the potential breeding pool, movement potential, and ecological niche of the organism. The movement potential of the organism refers to the maximum number of cells an individual can move during local migration and the species-specific entry cost for each cell. The ecological niche of the organism is determined by a vector of weights that are used in the calculation of an organism's fitness. These weights reflect which environmental variables are the chief determinants of fitness for a species.

1.4 Fitness function

The fitness function for an individual is a modification of Fisher's geometrical model of evolutionary adaptation (Fisher, 1930). Each organism has a phenotypic vector ($\mathbf{P}$), and high fitness corresponds to a close match between the phenotypic vector of organism and the environmental parameters ($\mathbf{E}$)

of the cell. Specifically, the logarithm of the fitness function is a weighted least-square function, where the weights are the species-specific parameters ($\mathbf{S}$) that control the species' niche. Thus, if the length of the vectors is Q, then the logarithm of the fitness for individual i of species j in cell k is:

$$\ln\left(\mathcal{F}_{ijk}\right) = -\sum_{q=1}^{Q} \mathbf{S}_q^{(j)} \left(\mathbf{P}_q^{(i)} - \mathbf{E}_q^{(k)}\right)^2. \tag{1.1}$$

The value of the fitness function directly gives the independent probability of survival of an organism of a particular species in a particular cell. In addition, the value of the fitness function also determines the ranking of the relative fitness of individuals within each cell during the competition phase. Note that the use of the Euclidean distance here assumes sphericity and identical units, following Fisher (1930). If the environmental vectors are not mutually independent, i.e. there is some co-variance, then this the Euclidean distance is not appropriate, and the Mahalanobis distance or some other approach is needed to combine the components (Waxman, 2006).

The phenotypic vector of an organism is inherited under the following model:

$$\mathbf{P}_q^{(i)} = \frac{\mathbf{P}_q^{(m)} + \mathbf{P}_q^{(f)}}{2} + Norm\left[0, \sigma = \sqrt{0.5}\right] \tag{1.2}$$

The inheritance of this fitness-determining phenotype is independent of the inheritance of the genealogies of the neutral loci that are tracked during the simulation. Following inheritance, the elements of the phenotypic vector are mutated using a species-specific probability distribution of mutational effects.

1.5 The simulation routine

Each generation or round of simulation consists of the following phases: 1. Landscape Configuration, 2. Reproduction, 3. Migration, 4. Survival, and 5. Competition. During the landscape configuration phase, the user-specified schedule of geographic parameters is used to alter the landscape. The carrying capacities, environmental parameters, cell entry costs, and long-distance dispersal

probabilities can be updated.

During the reproduction phase, all organisms of the same species within each cell mate randomly, producing a species-specific number of offspring. Offspring are assigned a gender with uniform random probability. The haploid allele of the mother is passed to the offspring. For each independent neutral diploid locus, the offspring inherits an allele at random from the diploid genotype of each of its parents. As noted above, the offspring's non-neutral phenotypic vector is inherited by combining elements from its parents' phenotypic vectors. Following reproduction, the parental generation organisms are removed from the simulation (generations are non-overlapping).

During the dispersal phase organisms move across the landscape subject to the constraints imposed by entry costs of cells.

During the survival phase, the fitness of every organism in its current cell is evaluated to give the probability of survival. Organisms that do not survive are removed from the simulation.

During the competition phase, the least-fit organisms are removed from each cell until the number of organisms in each cell is below the cell's carrying capacity.

The simulation proceeds for a pre-specified number of generations.. The user can determine when and how genealogies are sampled during the simulation. At the generation chosen, a random sample of organisms will be selected according to a sampling design which designates how many organisms are sampled from each cell. The genealogies for all of the loci for the selected individuals will be saved as trees in NEXUS files (Maddison et al., 1997). The tips of the tree are annotated with the XY-coordinates of the cell from which it was sampled. The internal nodes will be labelled with the XY-coordinates of the geographic position of the organism that held the most recent common ancestor gene copy. Thus, the resulting trees not only contain the phylogenetic history of the tracked loci, but also the full spatial or geographic history of the loci all the way back to the most-recent common ancestor of all the sampled alleles. The user can also specify that the total number of organisms of each species can be sampled at any point in the simulation. This occurrence sampling results in an ESRI ASCII raster grid format file for each species, where grid values represent the abundance of that species in each cell of the landscape.

1.6 Testing and Validation

We use unit testing (Zhu et al., 1997; Huizinga and Kolawa, 2007) to verify that all low-level program subcomponents (e.g., parsing of configuration files; individual organism movement, reproduction, survival; genealogy construction and serialization;) behave as expected.

Integrative testing and validation was provided by simulating data under a simple 4-island scenario, with each island exchanging migrants at an equal rate, and comparing the fixation index, F_{st}, calculated on sequences simulated on the resulting genealogies (using Seq-Gen, Rambaut and Grassly (1997)) to values predicted by the finite-island model (Nei et al., 1977).

We simulated data in `Ginkgo` under a demographic model consisting of a population subdivided into 4 demes exchanging migrants at an equal and constant rate, with all other elements of the simulation set to neutral or disabled altogether (e.g., environmental selection, competition, etc.). This corresponds to a Wright-Fisher population evolving under a finite-island model, and it has been shown that the fixation index, F_{st}, predicted under this model is (Nei, 1975):

$$F_{st} = \frac{1}{Nm\left(\frac{a}{a-1}\right)^2} \tag{1.3}$$

where: N is the size of each sub-population,
m is the proportion of migrants in each sub-population,
a is the number of islands.

While migration rates were fixed for any single simulation run, we selectively varied migration rates across runs ($0 < m < 1$), generating 100 replicate genealogies under each distinct migration rate.

Genealogies of the diploid locus of 25 random individuals were sampled from each population at various times during the simulation, yielding a series of 100-leaf genealogies. `Seq-Gen` (Rambaut and Grassly, 1997) was used to simulate a 1000-site nucleotide alignment on each of these genealogies, using a Jukes-Cantor finite-state model of sequence evolution with various mutation rates ($1e^{-8}, 5e^{-8}$, and $8e^{-8}$ per-site per-generation). F_{st} statistics were calculated on the simulated

sequences, and compared to those predicted by Equation 1.3 (Figure 1.1). As can be seen in Figure 1.1, while there is large variance (probably due to usage of a finite sites model and sampling error), in general the simulated data is consistent with the predictions of the 4-island incomplete subdivision model, as indicated by the best fit line ($R^2 = 0.9259$).

For comparative purposes we also simulated geneaologies using `ms` (Hudson, 2002), a program widely used in population genetic and phylogeographic studies for generating samples from a Wright-Fisher population, under the same demographic scenario (4-island model with migration rates that were equal and constant within each simulation, but varied across simulations). Sequences were simulated on each of the the genealogies using `Seq-Gen` (Rambaut and Grassly, 1997) under the same finite-state model and parameters. F_{st} statistics were calculated on the simulated sequences, and compared to those predicted by Equation 1.3. As can be seen in Figure 1.2, the data generated are very similar to that generated using `Ginkgo`, i.e., with large variance, but with the best fit line ($R^2 = 0.9216$) closely matching the perfect fit line.

1.7 Performance

When tracking 11 loci (10 diploid and one haploid), `Ginkgo` takes approximately 100 hours to complete 150,000 generation cycles with populations of 120,000 individuals evolving on a 50×50 cell landscape (running on a 3.3 GHz Intel Xeon machine). Memory usage peaks at approximately 2.75G, and stabilizes at around 2.25G. `Ginkgo` can also be built to track fewer loci, with dramatic improvements in performance. For example, if configured to track only one diploid and one haploid loci, the previous simulation would complete in approximately 10 hours, with memory usage not exceeding 0.5G.

Because `Ginkgo` is an agent-based simulator, its performance scales with the total number of organisms modeled. The upper-limit on the number of organisms can be specified (it is simply the sum of carrying capacities across all cells). Once the total number of organisms reaches its equilibrium number, computation speed per generation is constant. Computation time scales linearly with numbers of generations, provided that the cell carrying capacities do not change.

Memory usage behaves differently, however. Initially, Memory usage grows as a function of the product of the number of organisms and number of generations simulated. `Ginkgo` uses reference-

counting to track the genealogies of loci rather than organismal pedigrees. Thus, memory usage drops every time the genealogy of a tracked locus coalesces into a single individual. When this occurs, lineages without descendants are discarded and their memory is freed. Over time, the average memory usage tends to stabilize to an equilibrium level that is a function of total population size and the number of loci tracked.

1.8 Comparison with Similar Programs

The only other program currently available that provides for spatially-explicit forward-time individual-based simulation of genealogies is `DIM SUM` (Brown et al., 2009). `DIM SUM` uses a continuous landscape, with the positions of individuals tracked by real-valued coordinates (longitude and latitude), in contrast to the discrete cell-based landscape of `Ginkgo`, but uses a superimposed discrete grid to evaluate carrying-capacities and other population-level aspects of the simulation. The underlying genetic, demographic and spatial aspects of the simulation model of `DIM SUM` is much simpler than that of `Ginkgo`. For example, `Ginkgo` allows for simulation of multiple unlinked diploid loci, in addition to a single (maternally-inherited) haploid locus, evolving in sexually-reproducing populations of multiple species. In contrast, `DIM SUM` is limited to a single haploid locus evolving in an asexual-reproducing population of a single species; in effect, `DIM SUM` tracks (haploid) alleles rather than individuals. The spatial aspect of `Ginkgo` is also considerably more complex, with different species having different movement rates across the same landscape at different times in different places, and thus allowing for exploration of effect of distinct yet interacting ecologies on gene genealogy patterns. `DIM SUM`, on the other hand, has the same dispersal kernel for all individuals at all locations of the landscape. While `DIM SUM` offers carrying-capacity-based environmental regulation of organisms, `Ginkgo` allows for a complex multi-parameter spatio-temporally dynamic environmental selection regime. Apart from the long-term effects of selection on genealogies, this regime provides for far more realistic modelling of landscapes, as organismal movements can be restricted either by abstract movement rate limits or inhospitable environments, or both.

Despite the greater simulation model complexity, due in part to being implemented in C++ as well as other optimizations, `Ginkgo` runs much faster than `DIM SUM`. For example, a population of a 1000 individuals evolving on a 10×10 grid takes 22.56 minutes to complete 10,000 generations

under DIM SUM on a 3.3G machine. In comparison, running a similar-sized scenario under Ginkgo on the same machine (but tracking 10 diploid and one haploid loci in a sexual population instead of a single haploid locus in an asexual population) takes just 4.16 minutes to complete 10,000 generations.

While DIM SUM is much less complex and efficient in comparison to Ginkgo, it does allow for simulation of individuals on a truly continuous landscape, which can only be approximated in Ginkgo through the use of very fine-grained cells. In addition, the dispersal kernels used by DIM SUM may be easier to estimate and/or calibrate using available empirical data, while, with the abstract movement cost system of Ginkgo, some preliminary exploration and trial-and-error may be required to approximate the conditions of some real-world systems.

SPLATCHE (Currat et al., 2004) is another simulator that generates genealogies in a spatially-explicit framework. It uses a hybrid forward-time/coalescent approach to simulating genealogies and sequences within an environmentally-heterogenous spatially-explicit framework. While the SPLATCHE simulation system is also orders of magnitude less complex than Ginkgo, lacking, for example the ability to effect multi-scale spatial-structuring, multi-species competitive interaction, complex environmental selection and conditioning, etc., its hybrid approach allows for very efficient generation of genealogies when some of these aspects are not required.

1.9 Future Plans

Our primary motivation in the development of Ginkgo was to provide the software infrastructure necessary to characterize the performance envelopes of current phylogeographic analysis methods both under ideal conditions as well as when their assumptions were selectively violated. Ginkgo can be used, for example, to assess the false positive and false negative rates of these methods in identifying the correct phylogeographic history responsible for generating a particular set of data when the data were generated from populations evolving in classic Wright-Fisher conditions, as well as failure thresholds as these conditions are distorted in controlled and quantifiable steps in terms of geographical sub-structuring, selection, etc.

The Ginkgo simulation model is extremely complex. This complexity is to provide flexibility, so that different models and methods can be assessed in different ways, with selective and controlled

experiments, isolating and focussing on differnent aspects, assumptions or weaknesses of these models. While there is nothing in principle prohibiting the full suite of `Ginkgo` features being used in any particular application, the complexity of interacting parameters and features may make it very difficult to interpret any results, or ensure that any such interpretation is not artifactual. Thus, for example, while the software itself allows for high-dimensional environmental selection to be used in conjunction with nested spatial structuring, stochastic migration, and multi-species competition simultaneously, it would not only be challenging to calibrate the simulation parameters realistically, but also it would difficult to have any confidence that the results of a method applied to this simulation data are not being skewed by some abstract artifactual interaction of these parameter settings that would not be encountered in the real world.

While it is tempting to consider using `Ginkgo` as the simulation engine generating samples from the prior in an Approximate Bayesian Computation (ABC; Beaumont et al. (2002); Bertorelle et al. (2010)) context, the current state of computational power precludes this application for all but the simplest of studies. However, `Ginkgo` can still be used in ABC framework as a discovery and validation tool, to identify and develop useful summary statistics as well as to assess their power with respect to the prior models as well as robustness as these model assumptions are violated. We consider this, in fact, to be one of the more important applications of `Ginkgo` following its primary purpose of method evaluation discussed above.

In terms of `Ginkgo` features, we plan to incorporate a speciation mechanism in the next version of the program. This functionality is greatly desirable as it would provide the tools to understand how micro-level mechanistic processes may modulate the processes of speciation in a geographical framework, and allows `Ginkgo` to be sued as a platform to explore lineage diversification in a geographical context.

1.10 Availability

`Ginkgo` is released under the GNU General Public License 3+. Pre-compiled binaries for some platforms as well as user documentation are available for download from http://phylo.bio.ku.edu/ginkgo/, while the source code is available from the public source code management repository at http://github.com/jeetsukumaran/Ginkgo.

1.11 Figures

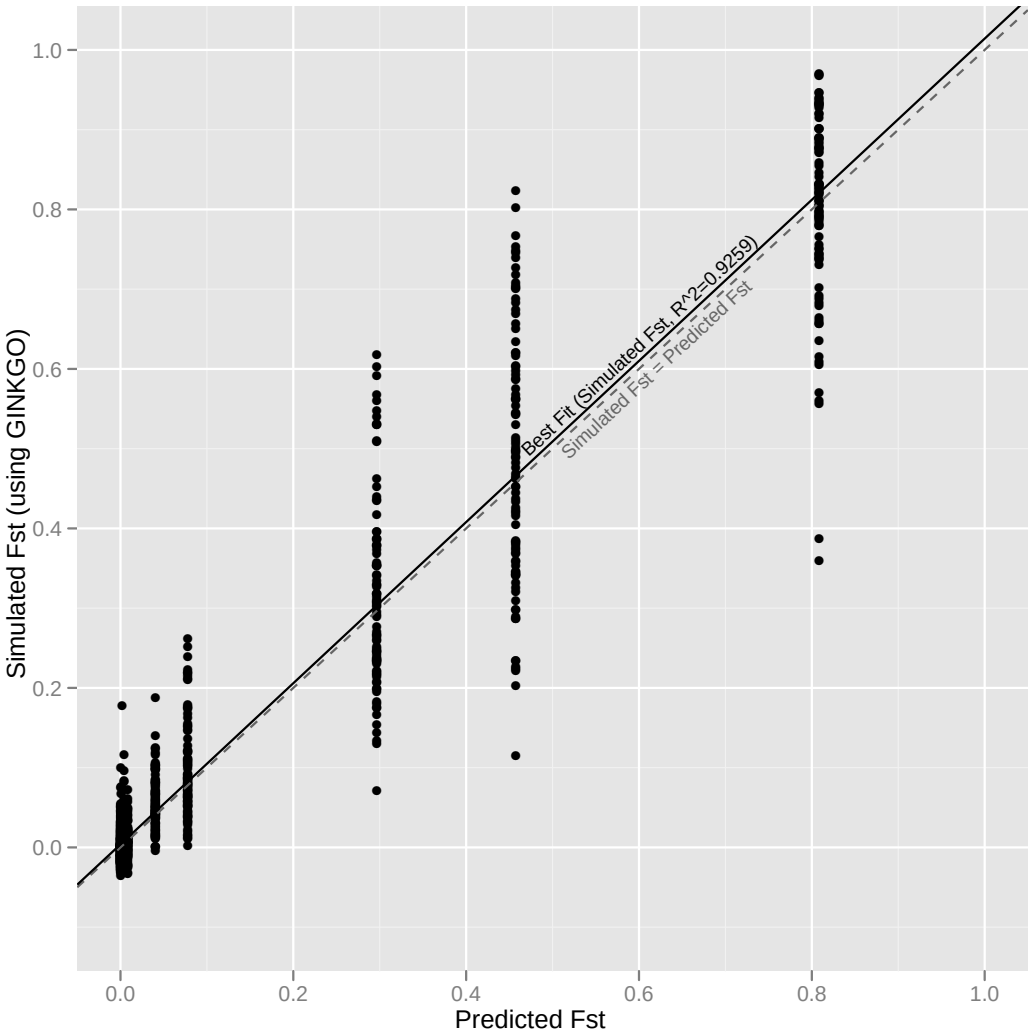


Figure 1.1: Integrative validation of the `Ginkgo` simulation engine. Y-axis represents fixation index (F_{st}) values, as calculated from sequence data simulated on genealogies produced by `Ginkgo` under a neutral 4-island model with migration, while X-axis represents corresponding values predicted by the n-island theory (Nei, 1975) for the same migration rate. Solid line shows best fit line (intercept $= 0.0037 \pm 0.0018$, slope $= 1.0103 \pm 0.0073$, $R^2 = 0.9259$), while dashed line shows the theoretical perfect fit line (i.e., intercept $= 0$ and slope $= 1$).

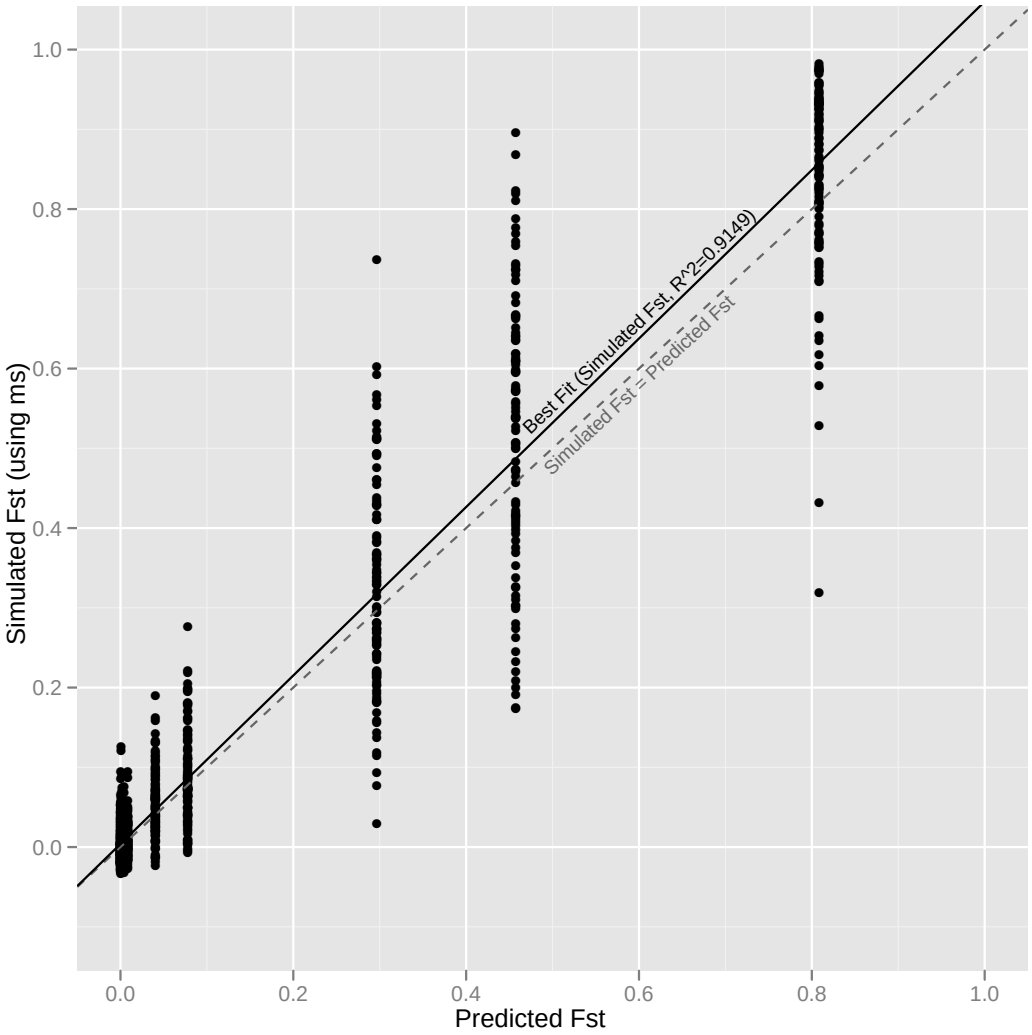


Figure 1.2: Simulation of the previous scenarios under `ms` for comparison purposes. Y-axis represents fixation index (F_{st}) values, as calculated from sequence data simulated on genealogies produced by `ms` under a neutral 4-island model with migration, while X-axis represents corresponding values predicted by the n-island theory (Nei, 1975) for the same migration rate. Solid line shows best fit line (intercept $= 0.0039 \pm 0.0019$, slope $= 1.0563 \pm 0.0077$, $R^2 = 0.9216$), while dashed line shows the theoretical perfect fit line (i.e., intercept $= 0$ and slope $= 1$).

Chapter 2

Evaluation of Performance of Approximate Bayesian Computation Approaches to Simultaneous Divergence Time Testing, with Comparisons to Full-Likelihood Approaches

2.1 Introduction

A number of biogeographic and phylogeographic hypotheses predict the simultaneous divergence of multiple groups of sister taxa. For example, the Kra Ecotone is a climatic and biotic transition zone located on the Thai-Malay Peninsula that marks the boundary between the mainland Indochinese biota to the north and the Sundaic biota to the south. This zone corresponds to the western boundary of the Sundaic region, just as Wallace's Line (Mayr, 1944; Simpson, 1977; Van Oosterzee, 1997), marking the transition from the Sundaic biota to an Australasian/Sahul biota, corresponds to the eastern boundary of this region. A wide range of biotic systems exhibit a shift in dominant

elements across the Kra Ecotone, including flowering plants, arthropods, reptiles, amphibians, fishes, mammals, birds etc. (Turner et al., 2001; Schulte et al., 2003; Michaux, 2010; Lourie and Vincent, 2004; Inger, 2005; How and Kitchener, 1997; Evans et al., 2003; Brown and Guttman, 2002; Baker et al., 1998; Inger, 1999). One class of traditional explanations for this pattern are climate-based, which posits the seasonal differences in precipitation (in particular, the minimum number of consecutive months without rain in the dry season) as the cause for the structure observed in the relationships of the biota between the regions on either side of the Kra Ecotone (Whitmore, 1987; Morley and Flenley, 1987). Woodruff (2003), however, suggests that marine highstands during the Neogene (Hall, 2001; Holloway and Hall, 1998) may have imposed barriers to gene flow between populations on either side of the ecotone, which have left their signatures in diversity patterns we see today. A clear and testable prediction of the Neogene marine highstand vicariance hypothesis is that pairs of sister taxa co-distributed across the Kra Ecotone would share the same divergence time. If climatic factors were the underlying cause for the patterns, on the other hand, there is no reason to suppose that the divergence times would be shared. Another example from Southeast Asia can be found in the Philippines. The eustatic lowering of sea-levels during Pleistocene hypothermals resulted in groups of smaller islands aggregating into "superislands", as the channels between them became exposed. The fragmentation of these Pleistocene Aggregate Island Complexes (PAIC's) of Greater Luzon, Greater Negros-Panay, and Greater Mindanao as the sea levels rose again after the Pleistocene has been suggested as the historical basis for inter-island relationships and patterns of endemism observed in various groups, such as mammals, amphibians and insects (Heaney et al., 2005; Brown and Guttman, 2002). A prediction that follows from this hypothesis is that sister taxa co-distributed across any of the within-PAIC component islands would share the same divergence time, dating back to the Pleistocene.

Ideally, these predictions of simultaneous divergence of co-distributed sister taxa could be tested by directly comparing divergence times of the splits in question from a time-calibrated phylogeny estimated on data collected from these systems. In many cases, however, the lack of suitable and reliable fossils to provide the calibration makes this approach unsatisfactory, due to large error or uncertainty that results from being forced to rely on a few, distant, external calibration points. Hickerson et al. (2006b) developed an Approximate Bayesian Computation (ABC) approach to testing this prediction of simultaneous divergence without using fossils, `msbayes`, and since then

this method has been applied in a number of different empirical studies: e.g. Leache et al. (2007); Daza et al. (2010); Barber and Klicka (2010).

Approximate Bayesian Computation is a Bayesian approach to estimating the posterior probability of a model (and parameters) **H** given the data **X** (Beaumont et al., 2002; Didelot et al., 2011; Beaumont, 2010) without calculating the likelihood. The calculation of the marginal likelihood of the data required to evaluate the posterior probability under Bayes Theorem is usually a high-dimensional integral and is difficult to compute. Thus, instead of evaluating the posterior probability directly, most full-likelihood Bayesian approaches sample the posterior probability distribution using Markov chain Monte Carlo (MCMC), importance sampling (IS), sequential Monte Carlo (SMC) or other approximation methods. While these approximation approaches avoid the often intractable calculation of the marginal likelihood of the data, they all still require calculation of the likelhood for every sample, and this is still a computationally expensive operation, even if tractable.

Approximate Bayesian Computation, in contrast, avoids the calculation of the likelihood by the use of a set of summary statistics. Samples are drawn (simulated) from the prior, and rejected based on whether or not the distance between the summary statistics calculated on the samples and those calculated on the observed data falls under some pre-specified threshold. The samples from the prior that are not rejected are accepted as samples from the posterior, and the posterior probability of any particular parameter value is given by its weight or frequency of representation in the set of accepted samples.

The Approximate Baysian approach of Hickerson et al. (2006b) estimates the posterior probability of the numbers of divergence times under a hierarchical model (Figure 2.1). Specifically, their `msbayes` approach estimates the posterior probability of ψ, which indexes the number of distinct divergence times across the Y pairs of co-distributed taxa, $\psi \in \{1, 2, \ldots\ Y\}$, integrating over other parameters including the mutation rate, actual divergence times, and, optionally, post-vicariance migration between daughter populations.

Hickerson et al. (2006a) investigated numerous summary statistics for use in the problem of divergence time estimation, and based on the results of this study, a preliminary set of these were adopted for use in `msbayes` (Hickerson et al., 2006b). Various other summary statistics were added to the implementation of the `msbayes` pipeline, including Wakeley's ψ, which Huang et al. (2011)

reported functioned well in distinguishing migration from isolation.

While the initial work of Hickerson et al. (2006b) and the later work of Huang et al. (2011) employed thorough tests to evaluate the effectiveness and power of the `msbayes` approach and its associated summary statistics, in all cases the testing was carried out with simulation models that were identical to the estimation model. As such, we do not understand how the `msbayes` approach behaves when dealing with data that does not conform to the model and its underlying coalescent assumptions. For example, the ancestral populations as well as the daughter subpopulations of all taxa are assumed to be pan-mictic with no substructuring, due to the Wright-Fisher premise of the coalescent. However, in many of the applications of this approach discussed above, as well as in many real-world populations, this is plainly false.

Furthermore, while the `msbayes` model does allow for incomplete isolation following the vicariance, i.e., migration or gene flow between the daughter populations, its performance has only been evaluated in a single simulation-based study (Huang et al., 2011). As previously noted, employed a simulation model that was identical to the estimation model. This does not allow for exploration of the method's behavior as assumptions of the model are violated.

Another issue is the distinction between parameter estimation and model selection using Approximate Bayesian Computation. Most work in the development of ABC methods, as well as investigations into the approach's effectiveness and robustness, has been in the context of parameter estimation in population genetics (e.g. Beaumont, 2010). Model selection, which is the objective here, in contrast, has only recently received attention, and numerous issues have been identified (Robert et al., 2011; Marin et al., 2011). For example, even if the summary statistics are sufficient, the posterior probability of one model relative to another maybe be incorrect by an arbitrary factor (Marin et al., 2011). As such, it would also be very useful to understand the performance of an ABC approach to model selection as given by `msbayes` in relation to a full-likelihood approach to model selection in phylogeography. An example of a full-likelihood method that might be used would be `IMa2` (Hey, 2010). `IMa2` uses a full-likelihood Bayesian Markov chain Mont Carlo approach to estimate population size, migration, and splitting time parameters for an ancestral population that splits into two or more daughter populations with possible post-vicariance gene flow (Figure 2.2).

If the mutation rates are known (and are constant), then the `IMa2` approach can be used to estimate the posterior probability of simultaneous divergence of multiple co-distributed taxon pairs.

This is because the IMa2 estimates of the divergence time are in units of mutation, and if the mutation rates of different taxa are known, even to an arbitrary common constant factor, the estimates of divergence times can be scaled, the results can then be related across multiple independent runs for different taxa. Generally, the mutation rate is an unknown parameter, and thus using IMa2 to test for simultaneous divergence is not viable approach for most empirical studies. However, in simulation-based studies the mutation rate is generally unambiguously known, and, furthermore, can be fixed to be equal across different taxa. This is key to being able to relate the results across the different taxa, and thus use IMa2 as a full-likelihood Bayesian approach to test for simultaneous divergence. This approach does still require some approximation (specifically, the binning of divergence times) due to the precision by which IMa2 reports its results. Nonetheless, while less than ideal, using IMa2 in this way does allow for some assessment of how a full-likelhood Bayesian model selection method may perform in contrast to an Approximate Bayesian Computation approach.

This study will characterize the relative performance of the msbayes Approximate Bayesian Computation and IMa2 full-likelihood Bayesian approaches to simultaneous divergence time testing under conditions that range from full conformance to the estimation model assumptions to controlled and selective violation of these assumptions:

1. The *baseline* cases, where simulated conditions approximate as nearly as possible the assumptions of the coalescent ancestral process that underly the estimation methods.

2. Cases with post-vicariance *migration*, where isolation between daughter subpopulations after the split is incomplete.

3. Cases with *substructuring*, where within-population substructuring (in both the ancestral as well as daughter populations) is strong enough to result in deviations from the pan-mictic assumptions of the coalescent.

This study will characterize each method's performance in terms of of its power to detect non-simultaneous divergence under the baseline configuration, as well as different levels of incomplete isolation and within-population substructuring. Forward-time simulations in a spatially-explicit framework will be used to generate the data used in the study, thus allowing for the assessment of the robustness of these coalescent-based phylogeographic model selection to detecting false patterns

when confronted with data with inherent spatial relationships (as described in, for example, Irwin, 2002).

2.2 Methods

2.2.1 Phylogeographic Simulation

Design

The core experimental design consisted of a two-species system, S_A and S_B, each with independent vicariance histories (Figure 2.3). The ancestral population of the first species, $S_A^{(anc)}$, split into two daughter subpopulations, $S_A^{(1)}$ and $S_A^{(2)}$, at time $t = T_A$ generations. The ancestral population of the second species, $S_B^{(anc)}$, in contrast, split into two daughter subpopulations, $S_B^{(1)}$ and $S_B^{(2)}$, at time $t = T_B$ generations. The difference in two divergence times, $\Delta T = T_A - T_B$, thus represents the separation in time between the divergence events in units of generations. Simulations were run under a range of divergence-time separation models. The simulation models in which $\Delta T = 0$ represented the cases of true simultaneous divergence, while the simulation models where $\Delta T > 0$ represented the cases of true non-simultaneous divergence. In the analyses and discussions that follow, the meta-parameter, ψ, will be used to index the number of distinct divergence times in the simulation, with $\psi = 1$ representing the single divergence time model and $\psi = 2$ representing the multiple divergence time model.

The simulations were carried out under three classes of conditions or configurations:

1. The baseline conditions were ones with both complete post-vicariance isolation between the daughter subpopulations, as well as unsubstructured populations. That is, following the vicariance event splitting the ancestral population $S_A^{(anc)}$ at T_A, no migrants were exchanged between $S_A^{(1)}$ and $S_A^{(2)}$, while following the vicariance event splitting the ancestral population $S_B^{(anc)}$ at T_B, no migrants were exchanged between $S_B^{(1)}$ and $S_B^{(2)}$. In addition, movement of individuals *within* each of the subpopulations was allowed to be as unrestricted as possible, so as to approach as nearly as possible the pan-mictic reproduction assumptions of a Wright-Fisher population.

2. Under the incomplete isolation conditions, some degree of gene flow persisted between $S_A^{(1)}$ and

$S_A^{(2)}$ after the vicariance event at T_A; that is, isolation between the daughter subpopulations of species S_A was not complete. Isolation between the daughter subpopulations of S_B, on the other hand, remained complete. All populations were unsubstructured, i.e., approaching Wright-Fisher pan-mixia.

3. Under the within-population structuring conditions, movement of individuals *within* each of the subpopulations ($S_A^{(1)}$, $S_A^{(2)}$, $S_B^{(1)}$ and $S_B^{(2)}$), were restricted, thereby introducing effects of isolation by distance by varying degrees. Isolation between the daughter populations following the respective vicariance events was complete in all cases.

Implementation

Forward-Time Simulations (`Ginkgo`) The forward-time simulations were carried out using `Ginkgo`, which generates data for multilocus diploid sexual (dioceous) individuals in a spatially-explicit framework (Sukumaran and Holder, 2011). An initial set of simulations of 10 replicates under the baseline configuration was carried out and analyzed to explore the parameter space and identify regions of interest to be investigated in further detail in subsequent simulations. Following this, an additional 10 replicates each under the three levels of post-vicariance gene flow and the three levels of within-population structuring were also carried out.

A separate `Ginkgo` simulation was run for each species (S_A and S_B) for each replicate under each experimental configuration and combination of parameters. Each `Ginkgo` simulation was carried out using a landscape grid consisting of 7 rows and 14 columns. Two 5×5 regions were established within this grid, corresponding to the regions occupied by each daughter subpopulation. The carrying capacity for each of these regions were set such that the total number of organisms across both regions corresponded to different values of the daughter subpopulation size, N. Various values of N were explored in a pilot set of simulations, $N \in \{500, 1000, 2500, 10000\}$, to determine optimum values that satistified practical as well as theoretical criteria. In particular, population sizes had to be small enough to allow for the completion of large numbers of simulation replicates, yet large enough so as not to result in artifactual behavior. These pilot studies indicated that subpopulation sizes of as small as 2000 or greater resulted in reasonable and similar behavior (corroborated by the backward-time simulations; see below), as long as mutation rates were scaled accordingly. As

such, due to the efficiency in run times, in subsequent simulations daughter population sizes of $N = 2500$ were used. This corresponded to a carrying capacity of 100 individuals per cell, and an ancestral population size of 5000. Outside the subpopulation regions, the carrying capacity was set to 0, such that no organism would be able to survive a generation there.

`Ginkgo` uses an artificial economy of "movement costs" for regulating the movement and interaction of individuals within its spatially-explicit framework (Sukumaran and Holder, 2011). Each individual organism in the simulations had a Poisson-distributed movement "budget" with a mean of 3. For the baseline cases, the entry costs for each cell *within* each subpopulation's region was set to the minimum value of 1. Previous work (Sukumaran and Holder, 2011) has indicated that this allows for within-population movement that approximates a Wright-Fisher population sufficiently well enough such that results can be predicted by both classical population genetics as well as coalescent theory. Movement outside the subpopulation regions was restricted by imposing a cell entry cost of 99.

The two subpopulation regions were initially placed adjacent to each other, with no restriction of movement from one region to another, such that both subpopulations effectively functioned as a single continuous population. At $t = 10N$ generations after the start of the simulation (T_A), a vicariance event was simulated in S_A by separating its two subpopulation regions by a region of cells in which the entry costs were 99 and the carrying capacity was 0. This effectively prohibited any interaction or gene flow between the two subpopulation regions.

For true simultaneous divergence simulation configurations, the ancestral population of S_B diverged at the same time as the ancestral population of S_A in corresponding simulations. For non-simultaneous divergence configurations, the ancestral population of S_B diverged at $T_A + 4N, T_A + 8N, T_A+16N, T_A+32N$ generations in different experimental configurations. In all cases, divergence was carried out as described in the case for S_A, i.e., by introducing a region that posed a barrier both to movement and occupancy.

For the cases with incomplete post-vicariance isolation, the "stochastic long-distance" dispersal feature of `Ginkgo` was used to to established controlled gene flow. Three different levels of incomplete isolation were simulated, resulting in individual per generation migration rates of $m \in \{0.000025, 0.000067, 0.00400\}$, corresponding to low ($F_{st} = 0.8$), high ($F_{st} = 0.6$), and very high ($F_{st} = 0.02$) migration rates, respectively.

For the cases with within-population substructuring, the movement currency of individuals within each subpopulation remained the same, but the cell entry costs were increased. Cell entry costs of 6, 9, and 11, were used, to result in low, medium and high within-population substructuring respectively.

Multiple replicates ($n = 10$) were run for each distinct experimental configuration and combination of parameters. Each replicate had multiple sampling periods, in which 25 individuals were sampled at random from each of the daughter populations $S_A^{(1)}, S_A^{(2)}, S_B^{(1)}$, and $S_B^{(2)}$. For the initial set of simulations, a sampling period density was high, with samples taken every $2N$ generations after the second vicariance event, T_B. For subsequent simulations, a reduced sampling regime was used, with samples taken at $T_B + 4N$, $T_B + 8N$, $T_B + 16N$, and $T_B + 32N$ generations after the second vicariance event, T_A. Each sample resulting in two sets of genealogies per sampling period: a set of genealogies for the diploid loci of the 50 individuals in total sampled from $S_A^{(1)}$ and $S_A^{(2)}$, as well as a set of genealogies for the diploid loci of the 50 individuals in total sampled from $S_B^{(1)}$ and $S_B^{(2)}$. Both single-locus and multi-locus (L=5) samples were used. These pairs of sets of genealogies, one pair of sets per sampling period per replicate per distinct experimental configuration and combination of parameters, constituted the final output of the phylogeographic simulation phase of the study.

Backward-Time Simulations (`ms`) The backward-time simulations were carried out using `ms` (Hudson, 2002), which generates data under the coalescent for a haploid unisexual single locus system. Two separate simulations were run for each replicate of each distinct combination of parameters (population size and difference in divergence times), one for S_A and one for S_B. Each simulation produced a sample of 50 individuals from two populations of size N each, with 25 individuals sampled from each subpopulation. Each replicate thus resulted in two genealogies of 50 individuals each, with one genealogy relating 25 individuals in each daughter population of S_A, $S_A^{(1)}$ and $S_A^{(2)}$, and the other genealogy relating 25 individuals in each daughter population of S_B, $S_B^{(1)}$ and $S_B^{(2)}$. The daughter populations of S_A were set to merge looking backward in time at a range of times in the past, from $4N$ to $64N$, to replicate the corresponding sampling periods in the forward-time simulations. In the true simultaneous divergence configurations, the daughter populations of S_B merged at the same time as those of the corresponding S_A simulations. In the non-simultaneous

divergence configurations, the daughter subpopulations of S_B merged at $4N, 8N, 16N, 32N$, and $64N$ generations earlier (looking backward in time) than the corresponding S_A simulations, to simulate different separations in divergence times between the two species S_A and S_B. As the backward-time simulation performance was invariant with respect to population size, a greater range of population sizes was explored, $N \in \{500, 1000, 2500, 10000, 30000, 300000\}$. Only single-locus baseline configuration simulations were carried out using the backward-time simulations.

2.2.2 Sequence Data Alignment Simulation

All the methods being tested use alignments of nucleotide sequences as their basic input data. As such, 1000-character alignments of nucleotide sequences were simulated under each genealogy produced by the phylogeographic simulation phase. The genealogies, scaled in units of generations, were rescaled by the various per site per generation mutation rates to produce trees with edge lengths in units of expected numbers of substitutions. These trees were used as input to `Seq-Gen` (Rambaut and Grassly, 1997) under a HKY model Hasegawa et al. (1985) of character evolution with a transition-transversion ratio of 0.5, i.e. corresponding to a Jukes-Cantor model.

For the pilot studies, per site per generation mutation rates used were: $2e^{-8}, 2e^{-7}, 2.4e^{-7}$, and $7e^{-6}$. These mutation rates are higher than the mutation rates generally reported for eukaryotic organisms, which averages between $1e^{-10}$ to $1e^{-8}$ (Wakeley, 2009). However, the population sizes used in the simulation were small ($N = 2500$), and the higher mutation rates compensated for these to result in reasonable eukaryotic population parameter values, θ: $0.0002, 0.0020, 0.0024$, and 0.0700. These population parameter values are within the range reported by many empirical studies of eukaryotes (Wakeley, 2009). For the full analysis, alignments generated under per site per generation mutation rates of $2.4e^{-7}$ and $7e^{-6}$ were used, for two categories of population parameter values: "low" ($\theta = 0.0024$) and "high" ($\theta = 0.07$).

The pairs of alignments simulated on the pairs of sets of genealogies constituted the final output of the entire simulation phase, and were used as the primary input for the methods being evaluated.

2.2.3 Estimation Procedures

The simulation phases described above produced multiple pairs of sets of alignments which could be characterized by distinct combinations of the following parameters:

- N, the size of the daughter subpopulations; for the full analyses described below, $N = 2500$.
- ΔT, the actual separation or difference in divergence times between the two pairs of sister taxa, $T_B - T_A$, measured in units of N generations, where N was the subpopulation size; in production studies, $\Delta T \in \{0N, 4N, 8N, 16N, 32N\}$.
- m, per individual per generation post-vicariance migration rate between daughter populations S_{1-1} and S_{1-2}; $m \in \{0.000025, 0.000067, 0.00400\}$, corresponding to low ($F_{st} = 0.8$), high ($F_{st} = 0.6$), and very high ($F_{st} = 0.02$) migration rates, respectively.
- z, within-population substructuring (i.e., restriction of movement within each of the daughter populations, $S_A^{(1)}$, $S_A^{(2)}$, $S_B^{(1)}$ and $S_B^{(2)}$); these were `Ginkgo` specific values of 1, 6, 9, and 11 for the between cell "movement costs", and corresponded to minimal, low ($F_{st} = 0.02$), medium ($F_{st} = 0.6$), high ($F_{st} = 0.8$) degrees of substructuring, respectively.
- t_g, the time period in which the samples were taken, $t_g \in \{T_B + 4N, T_B + 8N, T_B + 16N, T_B + 32N\}$.
- L, the number of loci used (1 or 5).
- θ, the effective (per-site) population parameter of the alignment.

Each paired set formed the input data to the two approaches for simultaneous divergence testing being evaluated: an approximate Bayesian computation approach using `msbayes`, and a full-likelihood approach using `IMa2`.

Approximate Bayesian Computation Estimation of Support of Simultaneous Divergence

Data from the prior were simulated using the `msbayes` MTML pipeline (Huang et al., 2011), which uses a modified version of `ms` (Hudson, 2002), `msDQH`, to simulate data under the coalescent conditioned on a vicariance splitting an ancestral population. The hyperparameters of the prior distributions were specified based on estimates of the final data or on the known truth. The upper bound of the uniformly distributed prior on divergence time was set to twice the deepest divergence time across all the genealogies, scaled by the $4N$ (the corresponding lower bound is fixed at 0 by

the program). The lower bound of the uniformly distributed prior on the population parameter θ was set to 0, while the upper bound was set to twice the maximum θ estimated across all the alignments. Each analysis was run explictly excluding migration (i.e., 0 for the upper bound the uniformly distributed prior on the migration rate), as well as low (0.001) and high (0.01) migration rates. Fifteen million samples from the prior were generated, and this sample was used across all experiments.

Following Huang et al. (2011) the summary statistics used for the single-locus rejection sampling were:

- π, the mean pairwise differences between sequences.
- $\hat{\theta}_W$, Watterson's estimator of θ (Watterson, 1975).
- π_{net}, the difference between the mean pairwise differences of sequences *within* each daughter subpopulation and the mean pairwise differences of sequences *between* each daughter subpopulation.
- var($\pi - \theta_W$), the denominator of Tajima's D (Tajima, 1989), i.e. the variance of the difference between two difference estimates of the population mutation parameter θ.
- Wakeley's ψ (Wakeley, 1996).

In each analysis these summary statistics were calculated for the alignments of S_A and S_B independentally, for a total of 10 summary statistics per analysis. Following Huang et al. (2011), for multi-locus analyses, the mean of these statistics across the alignments of S_A and S_B were used.

Rejection sampling was set to accept only the 1000 samples out of the 15 million samples from the prior with summary statistics closest to that of the observed data. While the `msbayes` package does include a rejection component, it did not perform well with large, parallelized analyses. Instead, the rejection sampling procedure was carried out using `ABCToolbox` (Wegmann et al., 2010).

ABC estimates are known to be biased toward the prior when using large numbers of summary statistics, due to the "curse of dimensionality" (Beaumont et al., 2002; Beaumont, 2010; Leuenberger and Wegmann, 2010). Specifically, large numbers of summary statistics are required to obtain satistfactory performance when dealing with complex and/or parameter-rich analyses: the

summary statistics need to satisfy sufficiency, such that $\Pr(\mathbf{H}|\mathbf{X}) = \Pr(\mathbf{H}|S(\mathbf{X}))$. At the same time, however, it is very difficult to generate a sample from the prior that results in a summary statistic equal to or very close to the observed data. As such, a larger error tolerance is required to generate sufficient samples in the posterior if acceptance is based on only accepting samples from the prior within a particular distance from the observed. In the acceptence regime described here (and used by `msbayes`), a pre-specified proportion of samples from the prior are accepted, and thus there is an implicit tolerance given by the maximum summary statistic distance accepted into the posterior, and grows with increasing numbers of summary statistics as well. The problem with this greater tolerance is that samples from the prior that map to points in summary statistic space further away from the observed data are weighted equally to samples from the prior that map closer to the observed data. This results in a systematically biased estimate, with the bias toward the prior. Beaumont et al. (2002) presented a method to correct for this bias by applying local linear regression to parameter estimates, which results in parameter estimates being weighted inversely proportionaly to the distance between their associated summary statistics and the the summary statistics calculated on the observed data. This method can only be applied to continuous values. To take advantage of this, `msbayes` makes use of an index statistic, Ω, which is given by the variance in divergence times across a simulated sample divided by the mean. This statistic ranges from 0 (if all the divergence times in the sample are equal) to arbitrary large positive values. Following the rejection step, Ω is calculated for every sample in the posterior, and then weighted by local linear regression to correct the bias toward the posterior. Hickerson et al. (2006b) suggest that an Ω value 0.01 be used as a threshold for concluding simultaneous divergence. This approach, i.e., the determination of simultaneous divergence based on a meta-summary statistic that considers the ratio of variance of divergence times to the mean, not only allows for the application of local linear regression to correct the bias toward the prior inherent in ABC estimation with a large set of summary statistics, but also allows for a tolerance in how close different divergence times can be and still be considered simultaneous. In this study, we will assess the use of the local linear regression mode and mean of Ω as indicator statistics of the preferred or estimated divergence time schedule model (single/simultaneous vs. multiple/non-simultaneous), in addition to the raw estimate of the approximate posteiror probability (as given by the proportion of samples in the posterior simulated under that model).

Full-Likelihood Estimation of Support for Simultaneous Divergence

For the full-likelihood estimation of support for simultaneous divergence, `IMa2` was run independentally on each set of alignments for each species. Priors on the divergence time, population sizes and migration rate parameters were uniformly distributed, with the upper bounds set to the twice the maximum estimate across all the simulated data. A total of four chains was used in each analysis (one cold, and three heated). Pilot studies were run to determined suitable numbers of steps and sampling frequencies. It was found that a burn-in of 10,000 steps, and followed by 10,000 samples from the posterior with 100 steps between each sample was sufficient to obtain convergence, as determined by inspection of likelihood sample plots in `Tracer` (Rambaut A., 2007), as well as consistent results across multiple runs.

The results of an `IMa2` analysis—population sizes, migration rates, and, of primary interest in the context of this study, population splitting times—are all scaled by mutation rates. As the mutation rates were constant and equal between the two species in this study, however, estimates of divergence times of daughter populations from `IMa2` can be compared directly across the two species in the study. `IMa2` reports the probability density of the splitting time scaled by mutation rates of the two daughter populations in a two population system. This probability density is reported in bins of 0.06. The pilot studies showed that the precision of the MCMC procedure using the settings described above is insufficient to accurately compare divergence times at this level of precision, especially for low θ values. Larger bin sizes, on the order of 10.0 mutation units, were required for consistent results. The posterior probability of simultaneous divergence was computed by summing the probability mass in corresponding bins across the two independent runs and normalizing by the total probability mass.

2.3 Results

2.3.1 `msbayes`

Single Locus Baseline Cases

Figures 2.4 and 2.5 show the results of the `msbayes` analyses of single-locus data generated using forward-time simulations (`Ginkgo`). Each figure is divided into a series of strips, with the top strip

representing the cases where the true divergence model was $\psi = 1$ (i.e., $\Delta T = 0$), while subsequent strips showing cases where the true divergence model was in fact $\psi = 2$, with gradually increasing values of ΔT: $2N$ generations, $4N$ generations, $8N$ generations, etc. The sampling period is represented on the x-axis in units of N generations: i.e., the time after the second divergence (T_B) when the daughter populations of S_A and S_B were sampled. The posterior probability of a non-simultaneous divergence time schedule model (i.e., $\psi = 2$) is indicated on the y-axis of each strip. Thus, a high value on the y-axis for the *first* strip shows strong support for the *false* divergence model, while in subsequent strips a high value on the y-axis shows strong support for the *true* divergence model. Each black dot represents a single set of paired samples from a single replicate of the experimental configuration, for a total of 10 replicates at each sampling period. Each blue dot represents the median of samples from the ten replicates for that sampling period.

Figure 2.4 shows the the results of analyses of data generated under high θ values (0.07), while figure 2.5 shows the results of analyses of data generated under low θ values(0.0024).

From figure 2.4, it can be seen that for high θ values, the separation of divergence times must be at least $16N$ generations before the `msbayes` ABC approach is able to detect non-simultaneous divergence consistently and correctly. At $\Delta T = 8$ and $\Delta T = 4$, this approach generally (and incorrectly) prefers the single divergence time schedule model. Furthermore, it can be seen that the power to identify true non-simultaneous divergence is limited not only by a minimum time separation between the two divergence events, but also constrained to be within a maximum span of time elapsed since the divergence events: support for the correct multiple divergence model erodes as samples are taken later and later after the second divergence. This limited "window" of resolution, a minimum amount of time separation, and maximum time since the divergence events, essentially describes the performance envelope of this approach for this data. If the system is sampled from outside this window, then the `msbayes` ABC approach defaults to providing support for single divergence.

In contrast, at low levels of θ, as shown in Figure 2.5, the `msbayes` ABC approach is unable to identify non-simultaneous divergence at any point in time using the posterior probability of models, regardless of the amount of time separating the two divergences. At lower levels of θ, then, the performance envelope collapses: the approach unconditionally prefers a model of simultaneous divergence, regardless of the true processes that generated the data.

As discussed above, `msbayes` uses an index statistic, Ω, given by the variance in divergence times in a sample divided by the mean, to determine the divergence time schedule model instead of the posterior probability directly. This statistic is weighted using local linear regression to correct any bias toward the prior. Any result in which the weighted value of Ω (summarized from the posterior samples either using the mode or the mean) falls below a critical threshold of 0.01 is taken to be indicative of support for simultaneous divergence. Figures 2.6 and 2.8 show the proportion of replicates supporting simultaneous divergence when using the mode (figure 2.6) or mean (figure 2.8) of the local linear regression weighted Ω values calculated from the samples from the posterior for analyses carried out on simulations under baseline conditions and *high* theta values. Figures 2.7 and 2.9, on the other hand, show the same but for simulations under baseline conditions and *low* theta values. At high theta values, the performance of `msbayes` when using the mode of the local linear regression weighted Ω to determine support for non-simultaneous divergence (figure 2.6) is very similar to the performance when using the estimated approximate posterior probability directly (figure 2.4). At low theta values, in contrast, reliance on this statistic results non-simultaneous divergence being inferred for all replicates (figure 2.7). This is the opposite of the case with the estimate approximate posterior probability, which unconditionally supported *simultaneous* divergence when analyzing simulations with data generated under low theta values (figure 2.5). Using the mean of the local linear regression weighted Ω values to determine support for simultaneous divergence shows strikingly different behavior. At high theta values (figure 2.8), all replicates are inferred to have experienced non-simultaneous divergence. At low theta values (figure 2.9), using the mean of the local liner regression weighted values of Ω allows discrimination of non-simultaneoiusus divergence in some replicates with separation of divergence time of $32N$, and all replicates with a separation of divergence time of $64N$.

With Migration

Analyzed Under Models That Do Not Consider Migration Figures 2.10 through 2.12 show the posterior probability of non-simultaneous divergence when single-locus data generated under high values of θ and incomplete isolation were analyzed using the Approximate Bayesian Computation approaches using an estimation model that does not take into account migration. With low migration rates, shown in figure 2.10 the scatter in posterior probability is greatly increased with

respect to the baseline cases. However, despite the increased noise, within a performance envelope comparable to that of the baseline cases, the correct model is still preferred. With medium and high migration rates, shown in Figures 2.11 and 2.12 respectively, a multiple divergence time schedule model is strongly preferred, regardless of the of the actual divergence time separation or sampling period. All posterior probability results of analyses of simulations with data generated under low levels of θ are omitted due to, as with the baseline cases, complete and unconditional support for simultaneous divergence.

The results for incomplete post-vicariance isolation using the weighted modes of Ω under low θ regimes are shown in figures 2.13 through 2.15 for analyses under no assumption of migration. The pattern of performance degradation is similar here as well, with increased noise at low-levels of true migration and preference for multiple divergence times at high levels of true migration. Again, results for analyses of data generated under low θ regimes are omitted due to lack of signal as with the baseline cases, i.e., complete and unconditional support for simultaneous divergence.

The pattern of increased noise at low and medium levels of migration and loss of signal is also seen in analyses of the low θ regime simulations incorporating migration when using the weighted means of Ω (figures 2.16 through 2.18). Results for data generated under high θ regimes are ommitted due to lack of signal, with complete and unconditional support for non-simultaneous divergence as with the baseline cases.

Analyzed Under Models That Consider Migration Figures 2.19 through 2.21 show the posterior probability of simultaneous divergence when single-locus data generated under high values of θ and incomplete isolation were analyzed using the Approximate Bayesian Computation approaches using a low migration rate model. The performance is very similar to inference under a 0-migration rate model: increased noise when there were low levels of migration in the truth, to unconditional preference for the multiple divergence time schedule model when there were high levels of migration in the truth.

Figures 2.22 through 2.24 show the posterior probability of simultaneous divergence when single-locus data generated under high values of θ and incomplete isolation were analyzed using the Approximate Bayesian Computation approaches using a *high* migration rate model. At low levels of true migration, analyses under models that permit high migration tend to support the single

divergence model, regardless of the true separation in divergence times. Conversely, when true levels of migration are high, the analyses tend to support a multiple divergence time schedule model, regardless of the true separation in divergence times.

Figures 2.25 through 2.27 show the replicates in which the multiple divergence model was supported when using the weighted modes of Ω as an index of non-simultaneous divergence in a model that allows for post-vicariance migration. The pattern is the same as with using the direct posterior probability as an indicator: at low true levels of migration, the single divergence time schedule model is unconditionally preferred, while at high true levels of migration, the multiple divergence time schdule model is unconditionally preferred.

With Substructuring

Figures 2.28 through 2.30 shows the posterior probability of simultaneous divergence when single-locus data generated under high values of θ and subpopulation structuring were analyzed using the Approximate Bayesian Computation approaches. Generally, low and medium substructuring tend to increase the scatter in the results, as seen in figures 2.28 and 2.29 respectively. However, high levels of substructuring, as shown in figure 2.30, reduce the ability of the method to detect any non-simultaneous divergence, to the point where no preference for either divergence time schedule model is indicated. (Again, as above, all posterior probability results of analyses of simulations with data generated under low levels of θ are omitted due to, as with the baseline cases, complete and unconditional support for simultaneous divergence.)

When using the mode of the weighted Ω statistic to determine the divergence time schedule model, the increased scatter at low (figure 2.31) and medium (figure 2.32) levels of substructuring, and (general) loss of preference for either divergence time schedule model at high levels of substructuring (figure 2.33).

In contrast to the above, when using the mean of the weighted Ω statistic to determine the divergence time schedule model (for data generated under *low* theta regimes), we see a general trend toward support for the multiple divergence model (figures 2.34 through 2.36). This trend is most clearly seen under the highest levels of substructuring (figure 2.36), though it is also noticable at lower levels (figures 2.34 and 2.35) when compared to the baseline case (figure 2.9). Thus, instead of a lack of strong preference of either divergence time schedule model, as seen when relying on

the raw approximate estimated posterior probability or the mode of the weighted Ω index, relying on the mean of the weighted Ω results in preference for a multiple divergence time schedule model under substructuring regimes.

Using Multiple Loci

Figures 2.37 through **??** show the results of analyzing multiple locus cases across all the conditions discussed previously using `msbayes`. Specifically, `msbayes` was applied to data from 5 independent loci evolved under baseline conditions, incomplete post-vicariance isolation, and within-population substructuring. Under the baseline conditions (Figures 2.37), it appears that there is an initial improvement in power, in that non-simultaneous divergence is diagnosed with only $8N$ generations separating the two divergence events. At the same time, however, there is a loss of power with samples taken at $32N$. The results for the migration and the substructuring cases remain extremely poor and inconsistent, with little clear pattern (Figures 2.38 through **??**). In some cases there is strong support for the single divergence time model and at others for theultiple divergence time model, regardless of the true generating model.

2.3.2 `IMa2`

Single Locus Baseline Cases

Figures 2.43 shows the results of using `IMa2` to analyze low θ regime data under baseline cases, i.e., where all estimation model assumptions are met or at least closely approximated, and there is neither post-vicariance gene flow nor any significant within-population substructuring. Note that, due to the longer run times required, parameter sampling was sparser than that of the `msbayes` analyses. In general, the performance in terms of the chronological window within which non-simultaneous divergence is broadly comparable to that of `msbayes` when using the raw estimated approximate posterior probability, or the local linear regression weighted mode of Ω. Specifically, non-simultaneous divergence can be generally detected as long as a minimum of $16N$ generations separates the two divergence events.

With Migration

Figures 2.45 through 2.47 show the results of using `IMa2` to analyze low θ regime data under cases of incomplete post-vicariance isolation, where the estimation model does not account for migration, while figures 2.48 through 2.50 show the results when the estimation model does allow for migration. In all cases, there is a tendency toward preferring multiple divergence time schedule models, regardless of the true generating model.

With Substructuring

Figures 2.51 through 2.53 show the results of using `IMa2` to analyze low θ regime data under cases where there is low, medium, and high within-population substructuring, respectively. While any particular replicate results in preference for one model or another, there is no consistency in the result: under the same conditions, different replicates produce support for different models.

2.4 Discussion

2.4.1 Baseline Performance of `msbayes`

The analyses of the baseline simulations showed that the `msbayes` ABC approach is only able to correctly identify non-simultaneous divergence under a restricted range of conditions. Regardless of whether the raw posterior probability, the mode of the local linear regression weighted Ω statistic, or the mean of the local linear regression weighted Ω statistic was used as an indicator of the divergence time schedule model, non-simultaneous divergence was only able to be correctly detected when the time separating multiple distrinct divergence event was at least $16N$ generations. When using the raw posterior probability or the mode of the local linear regression weighted Ω statistic, the window for detection of non simultaneous divergence was also limited in terms of the amount of time elapsed since the divergence events: after about $30N$ generations or so, all signal was lost. Outside this window, i.e., if the time separating multiple distinct divergence evets was less than $16N$ generations, or the time elapsed since the divergences was more than $30N$ generations, the `msbayes` approach provides (usually very strong) support for a single divergence schedule time model, regardless of the actual, true generating model.

This performance envelope of `msbayes` under baseline conditions has some characteristics that are cause for concern. Firstly, the performance envelope is limited both in terms of the minimum amount of time that must separate two divergence events for them to be detectable as non-simultaneous divergence as well as in the maximum amount of time that can elapse since the divergence event. The existence of minimum threshold [1] for detection of non-simultaneous divergence is not as troublesome as the concurrent existence of an upper bound on this detection window in terms of the amount of time elapsed since the second divergence that is on the same scale and so close in value to the minium threshold. The resulting performance window is, in other words, relatively narrow, limited as it is by both a minium and maximum time that are on the other of $16N$ generations apart.

Perhaps more troubling than the narrow performance envelope *per se* is the fact that the performance envelope is in some sense limited by the very quantity or variable that the method is trying estimate (i.e., divergence time). This somewhat precludes the recourse of pre-determining the applicability of the method by using other information, simply because use of the method becomes redundant once the applicability is determined in this manner. That is, if an investigator is able to use other information to ascertain the suitability of application of `msbayes` (e.g., a fossil-calibrated time tree, which presupposes information on at least the relative mutation rate, if not the population parameter, θ), then the need to use `msbayes` is obviated, as the variables of interest have already been estimated.

Another issue of concern is that `msbayes` does not fail by producing inconclusive results, but rather by producing strong support or preference for one particular model, which may or may not be the correct model. This leaves the investigator in the unfortunate position of needing to treat a strong conclusive result in any particular empirical analysis with skepticism. Furthermore, in a biological or empirical context, the conclusion of simultaneous divergence is often the more interesting one (and often is, in fact, the hypothesis of interest, as discussed in the introduction). As shown, usage two out of the three possible indicators results in spurious strong support for this model when the data is sampled from outside the chronological or mutational constraints within which we can expect reasonable behavior from `msbayes`. Thus, perhaps ironically, it is when

[1] Determined to be $16N$ generations in the context of these simulations, though Oaks et al. (submitted) determined this threshold to be on the order of 4 million years in the context of an empirical study with 22 taxon pairs.

`msbayes` shows the support for the hypothesis of interest that the empirical investigator should be most cautious.

2.4.2 Effect of Migration and Substructuring on Performance of `msbayes`

Both migration and substructuring had strong and complex effects on the performance of `msbayes`, and these effects are summarized in Table 2.1.

Table 2.1: Effect of various factors on different indicator statistics used by `msbayes` to select divergence schedule models: posterior probability (raw estimated approximate posterior probability, as given by the representation of this model in the samples from the prior); mode and mean of the local linear regression weighted Ω statistic (given by the variance in divergence times divided by the mean). "Migration (Under-treated)" refers to cases when post-vicariance isolation is incomplete, but the estimation model does not allow for migration or does not allow for sufficient migration. "Migration (Treated)" refers to cases when post-vicariance isolation is incomplete, but the estimation model *does* take into account migration, and, furthermore allows for sufficient levels of migration. Key to cells: "(no effect)" = no strong or consistent effect; "1DT" = tendency toward preferring single divergence time schedule model, regardless of true divergence schedule model; "2DT" = tendency toward preferring multiple or non-simultaneous divergence time schedule model, regardless of true divergence schedule model; "No power" loss of ability of discriminate between competing models.

Indicator	High θ	Low θ	Migration (Under-treated)	Migration (Treated)	Substructure
Posterior probability	(no effect)	1DT	2DT	1DT	No power
Mode of Ω	(no effect)	1DT	2DT	1DT	No power
Mean of Ω	2DT	(no effect)	2DT	1DT	2DT

Table 2.1 shows the effects of different factors on the performance of the three different statistics used to select a divergence schedule model: the estimated approximate posterior probability of a model, given by the proportion of representation of that model in samples from the posterior; the mode of the local linear regression weighted values of Ω calculated on samples from the posterior; and the mean of the local linear regression weighted values Ω calculated on samples from the posterior. As noted above, at high θ regimes, the raw posterior probability and the mode of Ω perform adequately, but the mean of Ω indicates strong support for the multiple divergence time schedule model ("2DT" in table 2.1). Conversely, at low θ regimes, the mean of Ω performs adequately, but the raw posterior probability and the mode of Ω perform indicate strong support for the single divergence time schedule model ("1DT" in table 2.1).

When migration is not accounted for in the estimation model, or the estimation model does

not allow the correct level of migration (by placing too low a bound on the uniform prior on migration rates), all three statistics end up strongly supporting the multiple divergence schedule model, regardless of the true generating model. Coversely, when migration is incorporated into the estimation model at high levels (by placing an upper bound on the uniform prior on migration rates much higher than the true migration rate), all three statistics end up strongly supporting the single divergence schedule model, regardless of the true generating model. In fact, work on this aspect of `msbayes` by Huang et al. (2011) came to the same conclusion, i.e., that unless an informative, precise *and* accurate prior is placed on the migration rate, the `msbayes` analyses tends to be mislead. This situation parallels that of the chronological window of performance described above: the `msbayes` approach can only be safely used when it is augmented with external information (even as it attempts to estimate some of that some information), and when this information is not provided it does not fail with inconclusive results, but with positively misleading results. Huang et al. (2011) recommend a pre-analysis to determine the level of post-vicariance gene flow, and using these results to set informative and accurate priors on the actual `msbayes` analysis. For the present, this remains the only viable way to use `msbayes`.

Within-population substructuring had strong effect on the `msbayes` analysis. When relying on the raw estimated approximate posterior probability or the mode, the local linear regression corrected Ω to select a divergence schedule model, within population substructuring led to loss of power: the method was unable to conclude a preference for any particular model. This is a satistfactory way for a method to fail outside of its performance envelope. Unfortunately, when relying on the mean of the local linear regression corrected Ω to select a divergence schedule model, support for a multiple divergence time schedule model was indicated, regardless of the actual or true number of distinct divergence times.

2.4.3 Effect of Migration and Substructuring on Performance of the `msbayes` Summary Statistics

The performance of the `msbayes` ABC approach described here is based on the summary statistics used, and the limitations described above are attributable to the behavior of the summary statistics outside the performance envelope. A principle component analysis was carried out using the summary statistics calculated using the data generated using the forward time simulations,

under baseline conditions and high θ values. Figure 2.54 shows a plot of the components 1 and 2. The blue dots represent true simultaneous divergence (i.e., $\psi = 1, \Delta T = 0$), the green dots represent non-simultaneous divergence with $\Delta T = 4$ and $\Delta T = 8$, and the red dots represent non-simultaneous divergence with $\Delta T \geq 16$. As can be seen, the results are consistent with the full `msbayes` analyses: using these summary statistics, it is difficult to discriminate between true simultaneous divergence and $\Delta T = 4$ and $\Delta T = 8$, but higher differences in divergence times can be distinguished over a reasonable portion of parameter space. The portion of parameter space where even this discrimination collapses, toward the right-hand side of the plot, is where the samples were taken after $t = T_B + 30N$.

Table 2.2: Principle components analysis of summary statistics calculated on data across all simulations. See text for discussion.

Summary Statistic	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
pi.1	-0.412	-0.564	-0.179	0.166	
pi.2	-0.365	-0.389	-0.278		0.239
wattTheta.1		-0.134			
wattTheta.2					
pi.net.1	0.533	-0.584	0.292	0.411	-0.301
pi.net.2	0.605	-0.231	-0.474	-0.470	0.342
Psi.1	0.124	0.284	-0.719	0.572	-0.234
Psi.2	0.141	0.153	0.216	0.495	0.813
Proportion of Variance	0.7514058	0.09802093	0.08078942	0.05915855	0.005692997
Cumulative Proportion	0.7514058	0.84942672	0.93021614	0.98937469	0.995067684

Table 2.2, which summarizes the PCA on the observed summary statistics, shows that Component 1 explained over 75% of the variance, with the remaining components each explaining less that 1%. Of the individual summary statistics, π_{net} and Wakeley's ψ load positively on component 1, with π_{net} in particular loading very strongly (0.533 and 0.605 for the first and second taxon pairs respectively). Thus, π_{net} is the most important statistic in driving the results.

It is no surprise, then, that increasing within-population structuring results in non-informative results. As the within-population structuring increases, the mean pairwise distance between sequences within each population increases in relation to the mean pairwise distance between each population, and thus the mean difference between these values decreases, to the point where all signal is lost.

2.4.4 Multilocus Data

It has long been understood that increasing the number of independent loci increases the power of coalescent-based analyses (Kuhner et al., 2000; Nielsen, 2000; Carling and Brumfield, 2007). However, in this study, increasing the number of independent loci analyzed to five had in some cases, and somewhat counter-intuitively, an adverse effect on the results. There appears to be an initial increase in power, in terms of the divergence time separation resolution, from $16N$ to $8N$. However, the performance of analyses of data generated under non-baseline conditions actually degenerated. These are similar to conclusions drawn by Huang et al. (2011): when using the less than about 16 loci or so, the performance of `msbayes` was extremely poor relative to single locus analysis, and that the method requires more than 32 loci or so to benefit from multi-locus analysis. They conjecture that this degeneration in performance is due the additional rate hetereogneity introduced by multiple loci. However, in these simulations, actual subsitution rate was fixed to be equal and constant across all loci, so rate hetereogeneity per se was not a factor. One other possibility might be with the way that `msbayes` integrates the summary statistics from across the different loci: each summary statistic is calculated independentally for each locus, and the arithematic mean of the statistic across loci is taken as the summary statistic for the entire multilocus dataset. This approach not only discards much information, but also, due to the central limit theorem, actually changes the expected distribution of all the statistics to converge to a normal distribution. With extremely large datasets (i.e., on the order of 32 loci or so), enough information may remain with the data such that robust results might be recovered.

2.4.5 Comparison with Full-Likelihood Model Selection

The full-likelihood approach to model selection employed in this study did not perform much better than the approximate approach. Its performance envelope, in fact, on a broad scale, mirrored that of the approximate approach under all conditions. This finding must be mitigated or considered in the light of the fact that the full likelihood approach used in this study was constrained by the somewhat crude binning used to relate the analysis across the two independent runs. As noted, these bins were on the order 10 units of `IMa2` time, which corresponds to $16.67N$ generations. These large bins result in loss of discriminatory power, and probably one of the reasons that the

power of the full-likelihood approach was not clearly and significantly better than the approximate approach. (Finer binning did not result any shared probability mass for any particular time bin across pairs of runs, and thus could not be used. It is possible that if the `IMa2` MCMC were run out for much longer, it would attain the numerical stability and precision required to allow finer-scale binning of the time units. However, the run times for this would preclude large-scale evaluations of parameter space such as those conducted in this study.)

2.4.6 Implications for Empirical Studies

This study suggests that any investigator wishing to use `msbayes` to test the hypothesis of simultaneous divergence would need to:

1. Ensure that the divergence times of each the populations were sufficiently recent.
2. Ensure that the priors on migration rates are informative and accurate.
3. Ensure that there is no significant substructuring within any of the populations.
4. Recognize that "simultaneous" may actually indicate a fairly broad time span.

In this study, the "sufficient recent" was approximately $30N$ generations, while "simultaneous" was $16N$ generations. Obviously, what values constitutes "sufficiently recent", and the size of the possible time span that constitutes "simultaneous" might vary from system to system.

For example, the hypothesis marine highstands Miocene or Pliocene responsible for structuring the biota across the Kra peninsula, as described in the introduction, would imply divergence times of 24-13 MYA or 5.5-4.5 MYA. To apply `msbayes` to test the prediction of simultaneous divergence following from this hypothesis, an investigator would have to estimate the population sizes of the taxa being studied, as well as the generation time, to determine if the these period falls within the performance envelope. Furthermore, the investigator must also demonstrate that the none of the pairs of daughter populations being studied diverged considerably before the period of interest, based on, for example, the maximum age of the credibility interval for divergence times.

Empirical investigators will also have to accurately determine migration levels between their populations, using programs such as `migrate` (Beerli and Felsenstein, 2001), and set very tight priors on the migration rate based on these estimates. The current implementation of `msbayes`

does not allow for different migration priors on different groups, and this must be accommodated if this situation is encountered in empirical data.

Empirical investigators will have to demonstrate that the populations from which their samples are not excessively substructured. As with other factors, the quantitative or numerical definition of "excessive" might vary from study to study. Perhaps, to be conservative, some sort of AMOVA (Excoffier, 1995) or similar tests should be carried out to establish that there is no significant substructuring within the daughter populations, and there fore allow for legitimate application of `msbayes`.

All this indicates a broad suite of analyses that need to be carried out *before* `msbayes` can be applied, e These validation procedures were not carried out in any of the empirical applications of `msbayes` published thus Leache et al. (2007); Daza et al. (2010); Barber and Klicka (2010, e.g.). These, and other empirical studies, may need to be revisited in the light of the present study to establish whether or not their conclusions are artifacts of `msbayes` being confounded in one direction or another by any of the factors described here.

2.4.7 Final Conclusions

The co-divergence of multiple pairs of taxa is a common prediction of a large class of biogeographic and phylogeographic hypotheses. The most powerful and direct way to test this prediction, i.e., by directly comparing the ages of the appropriate splits on a fossil-calibrated ultrametric phylogeny, is often not feasible due to lack of suitable fossils. The `msbayes` approach, using Approximate Bayesian Computation, provided a promising alternative. Unfortunately, this approach is limited in a number of ways. It suffers from a narrow performance window coupled with the inability to determine whether or not the data were sampled from within that performance window. The fact that, in all but one class of conditions, `msbayes` produces positively misleading spurious results (instead of inconclusive results) when operating out that performance window is also very troubling.

Adding more loci does appear to improve the performance of `msbayes` with respect to the minimum amount of time separation needed to diagnose multiple divergence. The result of Huang et al. (2011) suggest that better precision and accuracy are obtained as the number of loci increase to 32, but the evaluation was carried out *within* the established performance window of `msbayes`. If it were in fact possible to determine whether the data indeed conformed to the method's performance

parameters (the time separating the divergence events, the time elapsed since the divergence events, or the mutation rates), then it would also be possible to apply more direct and exact methods to answer the question asked by the `msbayes` approach.

To improve `msbayes`, the most obvious approach would be to focus on its summary statistics. In principle, it should be possible to develop a set of summary statistics that perform better. The summary statistics used in the current implementation are population genetic summary statistics, and do not use phylogenetic tree or spatial information. It might be hoped that incorporating some of this information into summary statistics might yield better results. This view, however, is challenged by the fact that even the full-likelihood approach, `IMa2`, in general is constrained to the same performance envelope as the `msbayes` approximate approach in terms of the minimum difference in divergence time and maximum time elapsed. As noted, though, the full-likelihood method used here was crippled in terms of its power due to the poor resolution imposed by the binning of divergence times. Thus, the possibility remains open that different summary statistics might be able to produce a much more satisfactory Approximate Bayesian Computation approach.

An alternative might to focus on developing a full-likelihood approach to testing the prediction of simultaneous divergence. In the long run, the gains in computational flexibility and efficiency of the Approximae Bayesian Computaiton approach may not be sufficient to compensate for loss in power, especially as computation hardware improves.

2.5 Figures

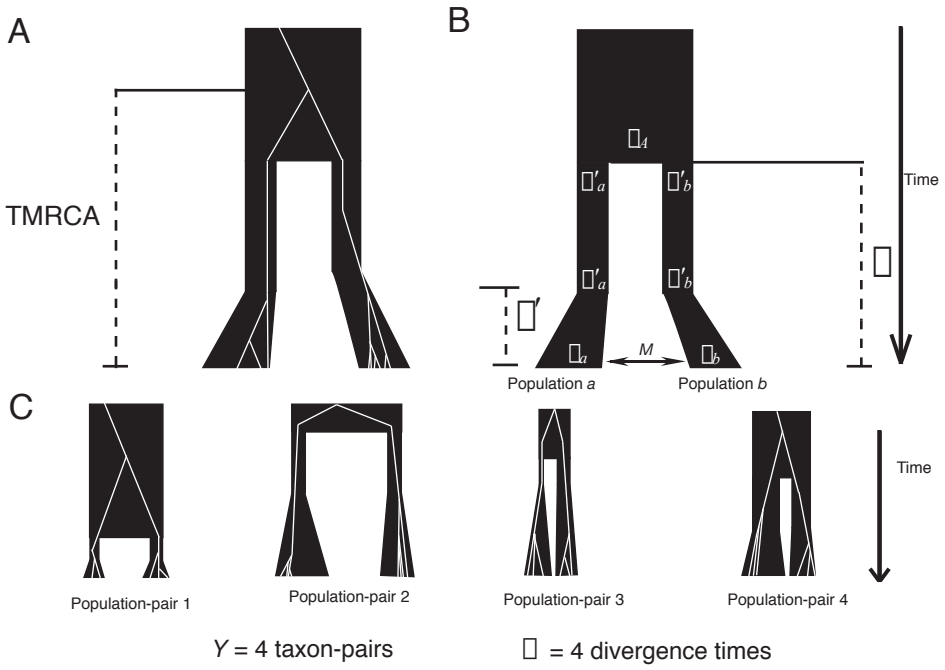


Figure 2.1: The hierarchical model used in `msbayes` to estimate the posterior probability of ψ, the number of distinct divergence times found in a sample of Y co-distributed taxon pairs (Hickerson et al., 2006b). (A) The white lines represent the genealogy coalescing within the containing black species or population tree. (B) The parameters of the model include the times of divergences for each population pair, here given as τ and τ'; the population parameters θ for the various populations and stages of populations, where $\theta = 4N\mu$, and N is the population size and μ is the mutation rate; the migration rate between the populations, here given as M; etc. Figure from Hickerson et al. (2006b)

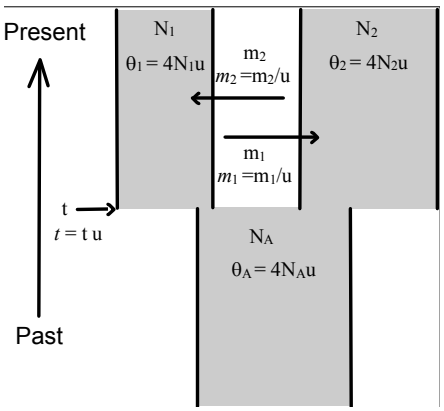


Figure 2.2: The isolation with migration (IM) model for two populations. Parameters include the population sizes (N_1, N_2, and N_A), per gene copy per generation migration rates (m_1, m_2), and population splitting time (t). All parameters are scaled by the neutral mutation rate (u) in the actual analysis. From Hey (2006).

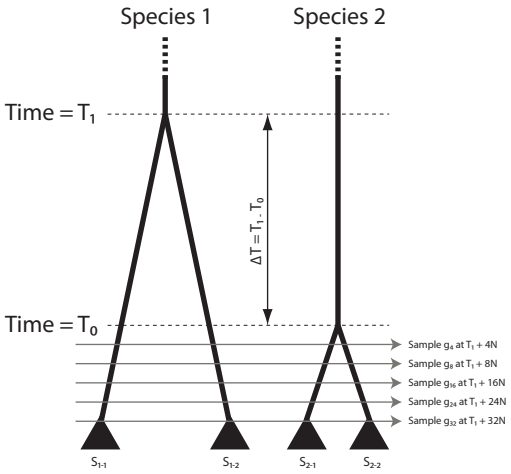


Figure 2.3: The fundamental experiment design that formed the core of all studies to assess the performance of simultaneous divergence time inference methods. The ancestral population of two species, $S_A^{(anc)}$ and $S_B^{(anc)}$ split into two daughter subpopulations at T_A and T_B. True simultaneous divergence are the cases where $T_A == T_B$; all other cases are non-simultaneous divergence. Samples are taken various intervals following the second divergence, and sequences simulated on these are used as the input for the simultaneous divergence time inference methods.

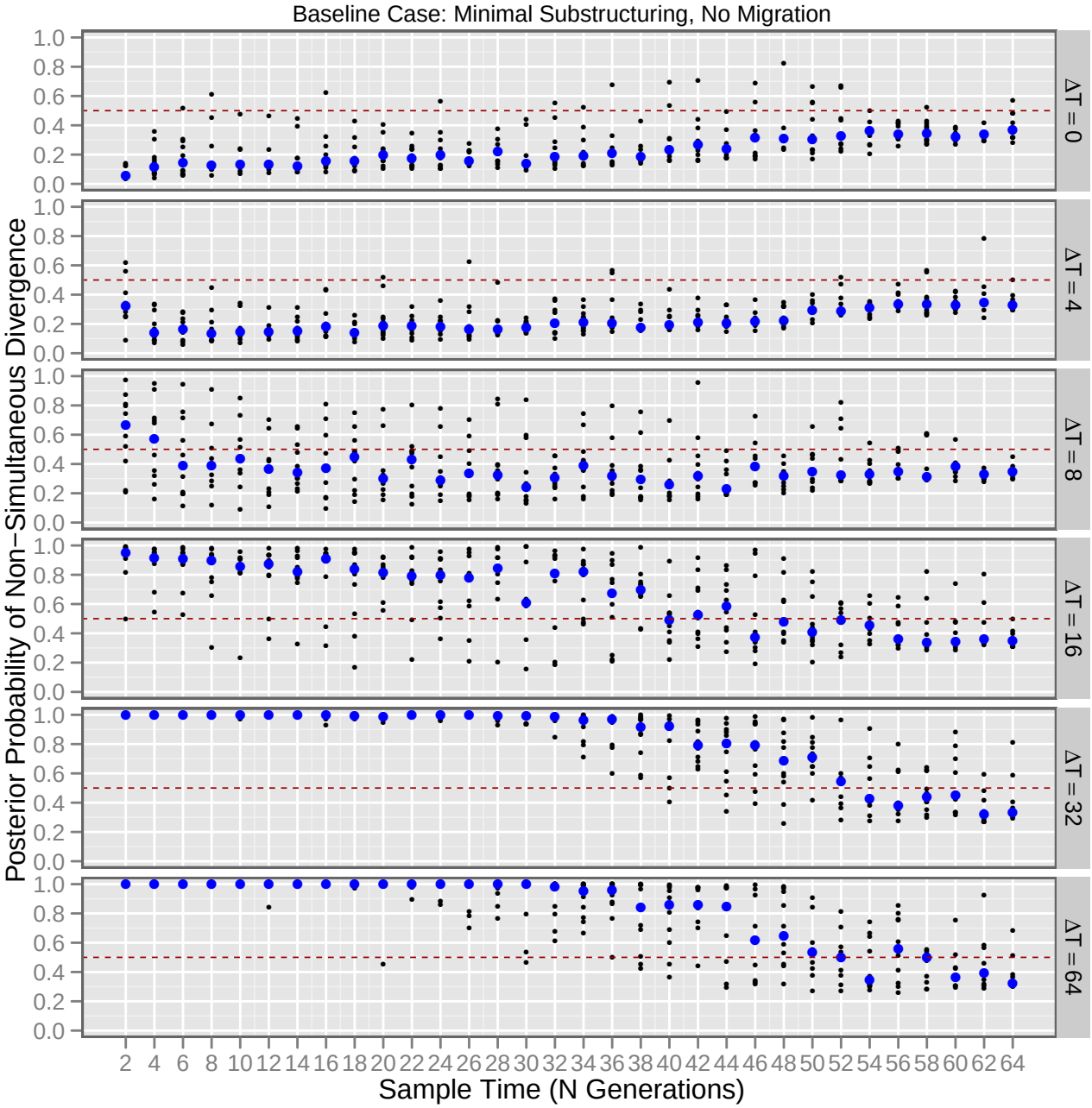


Figure 2.4: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under high theta values when analyzed using `msbayes`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. Y-axis on each strip indicates estimated approximate posterior probability of support for *multiple* divergences, while X-axis indicates period after the second vicariance event in which the sample was taken. Each strip is for 10 replicates of a simulation carried out at particular difference or separation of time between the two divergences. The top strip, $\Delta T = 0$, is when there is *no* difference in time between the two divergences, i.e., the case of true simultaneous divergence. High values on the Y-axis here indicate support for the *wrong* model. The remaining strips show increasingly larger differences in time between divergence events, and high values on the Y-axis thus indicate support for the *correct* model.

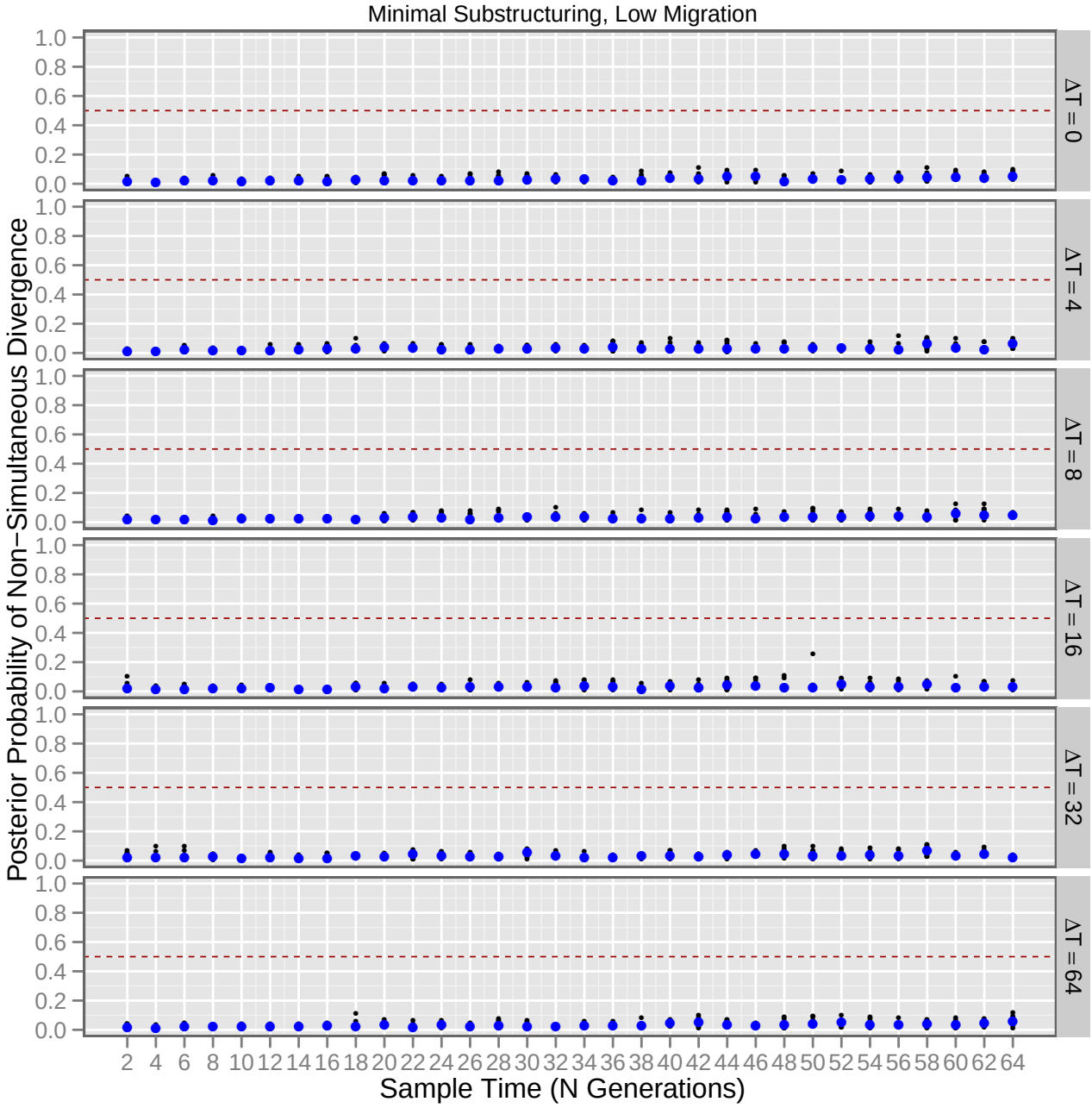


Figure 2.5: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under *low* theta values when analyzed using `msbayes`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.4 for details on interpreting the plots.

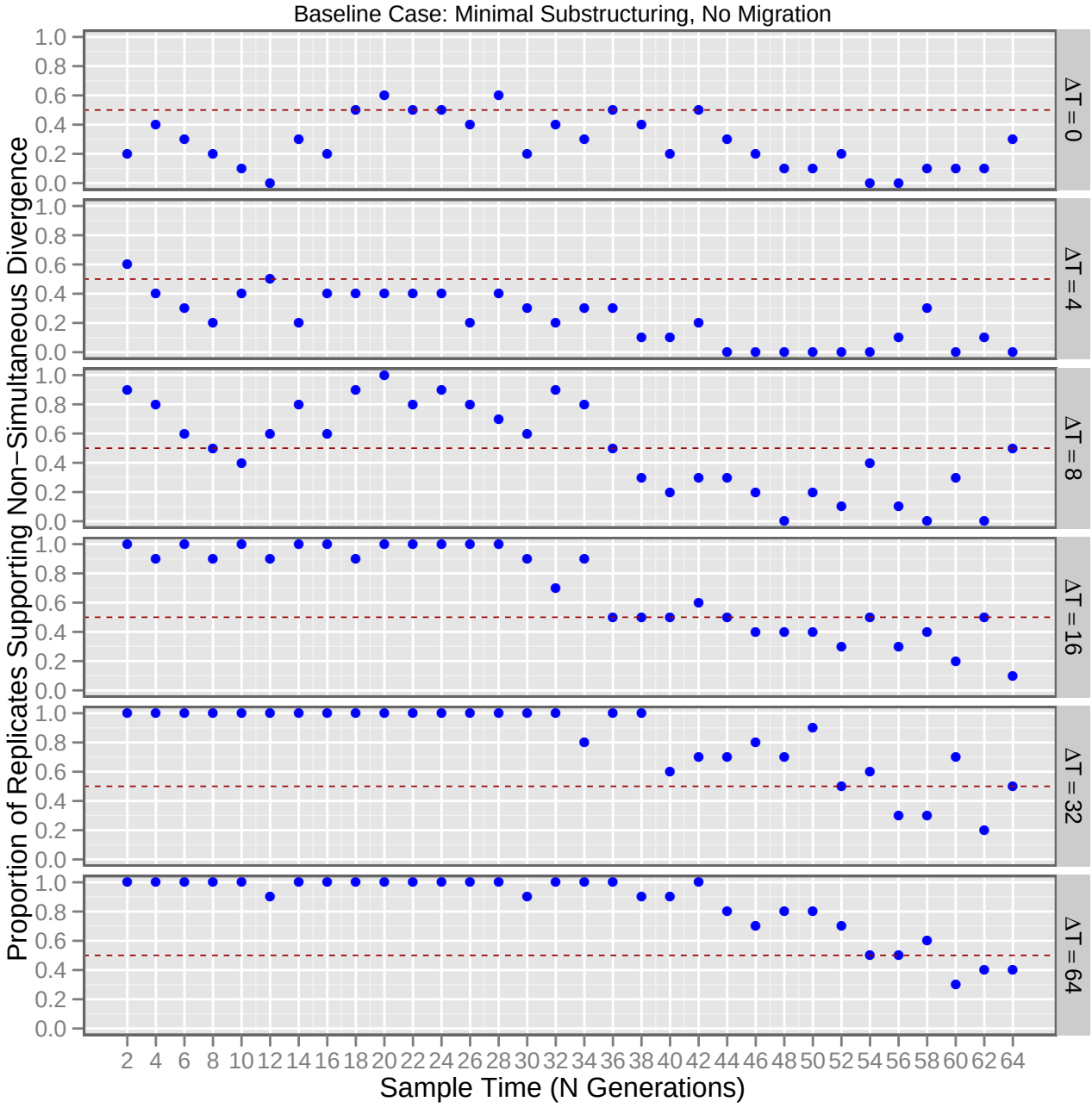


Figure 2.6: Proportion of replicates supporting non-simultaneous divergence of "baseline" forward-time simulations under *high* theta values when analyzed using `msbayes`, and using the *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. All assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.4 for details on interpreting the plots.

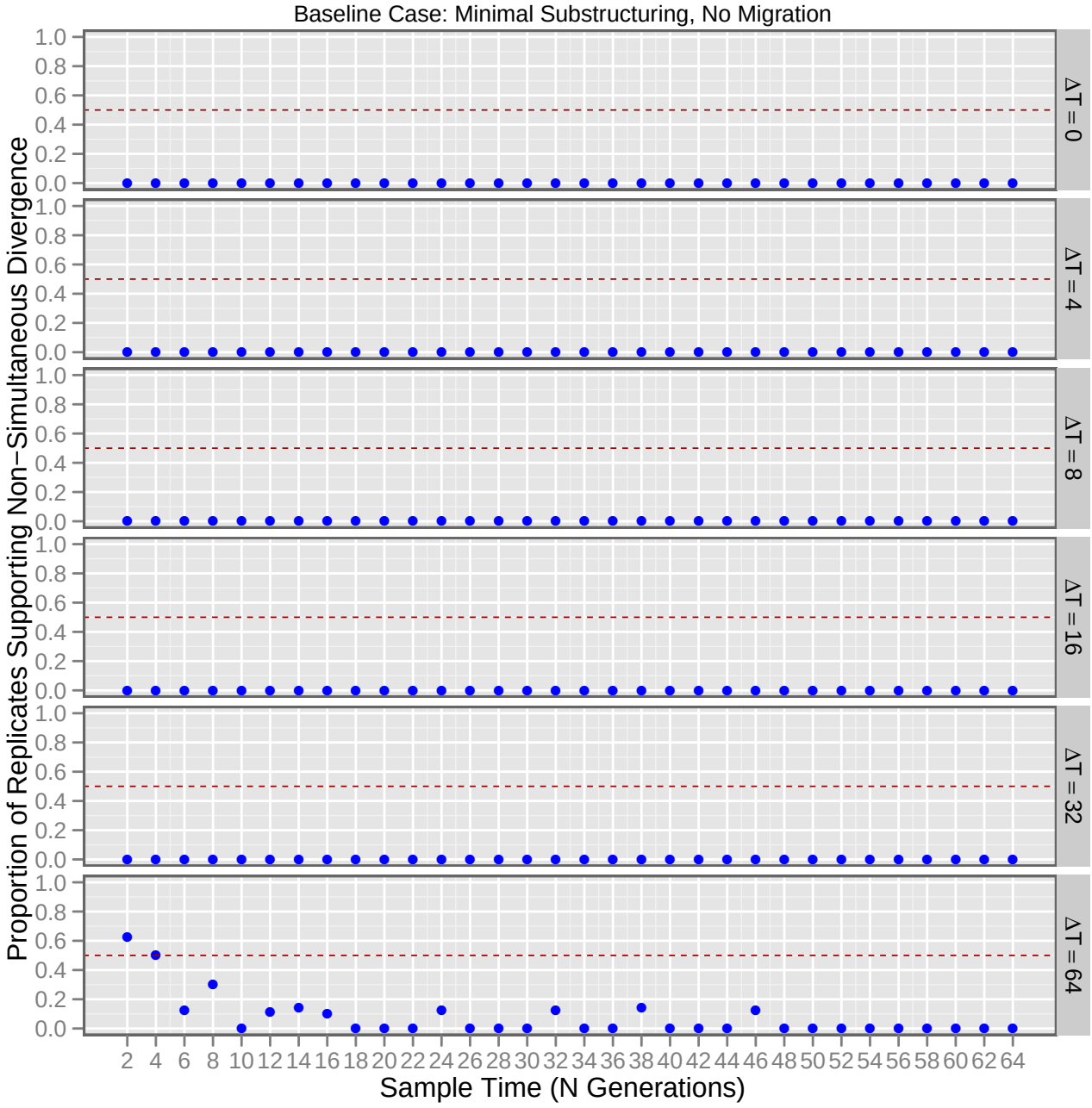


Figure 2.7: Proportion of replicates supporting non-simultaneous divergence of "baseline" forward-time simulations under *low* theta values when analyzed using `msbayes`, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. All assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.5 for details on interpreting the plots.

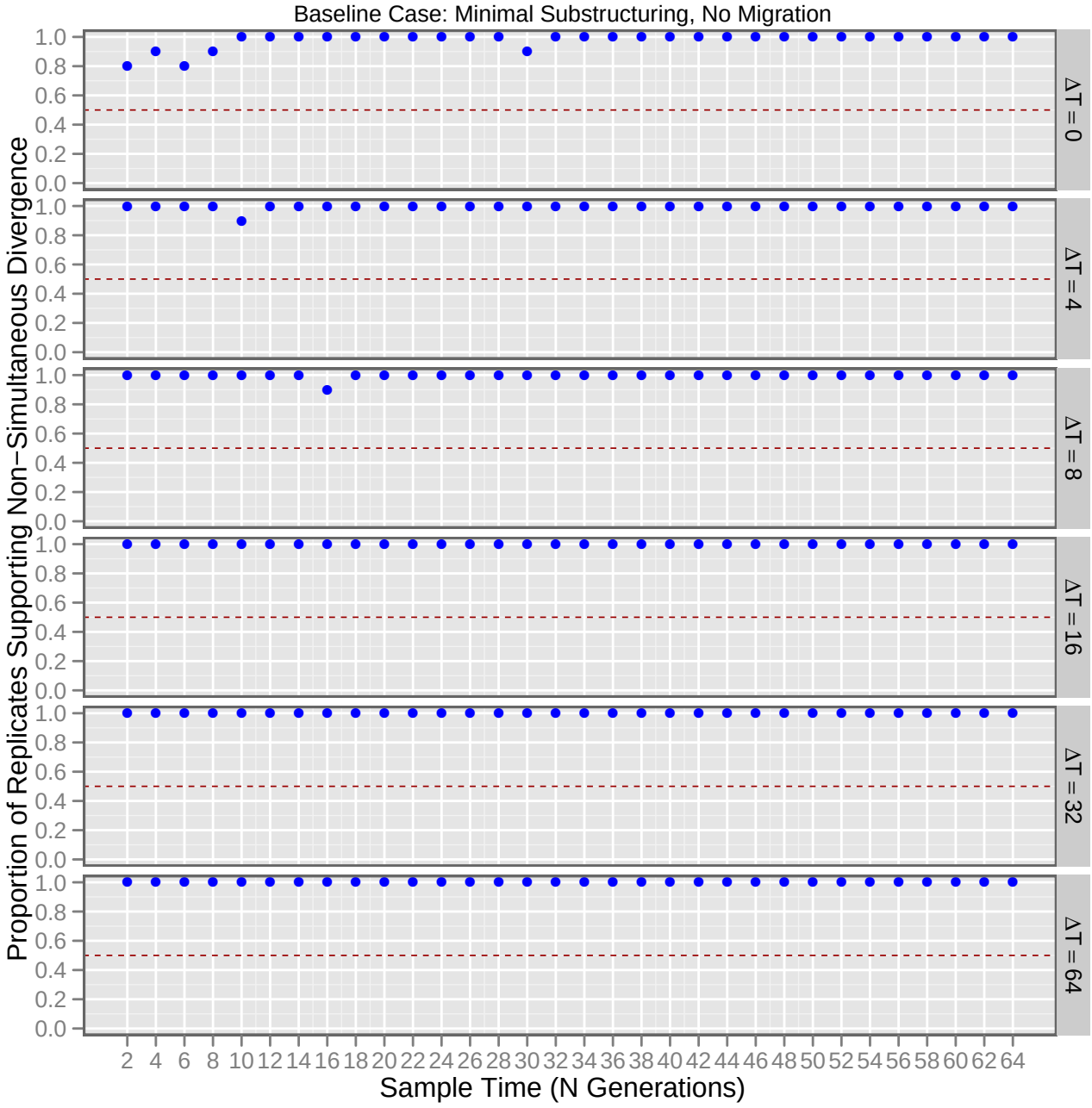


Figure 2.8: Proportion of replicates supporting non-simultaneous divergence of "baseline" forward-time simulations under *high* theta values when analyzed using `msbayes`, and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. All assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.4 for details on interpreting the plots.

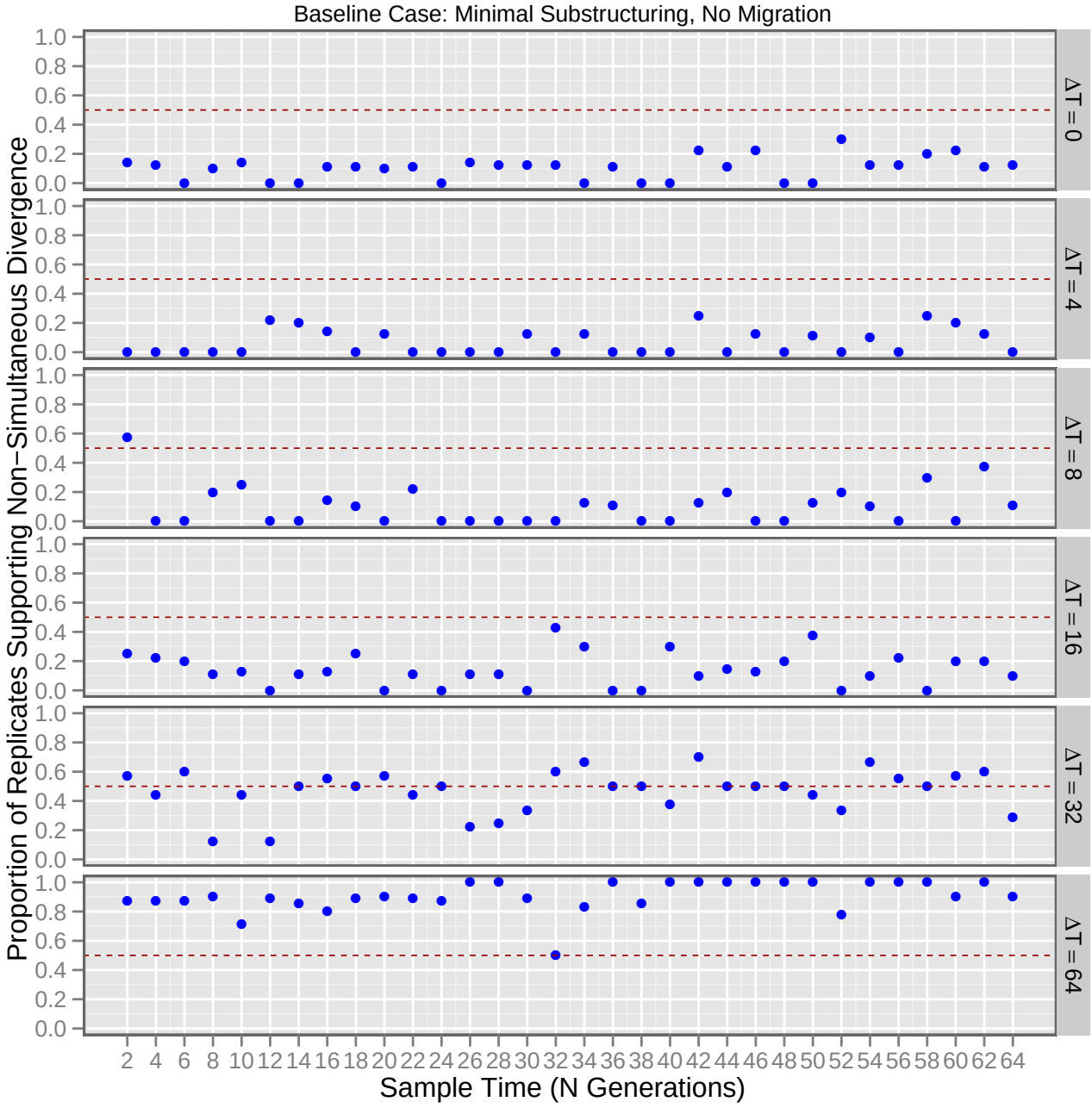


Figure 2.9: Proportion of replicates supporting non-simultaneous divergence of "baseline" forward-time simulations under *low* theta values when analyzed using `msbayes`, and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. All assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.5 for details on interpreting the plots.

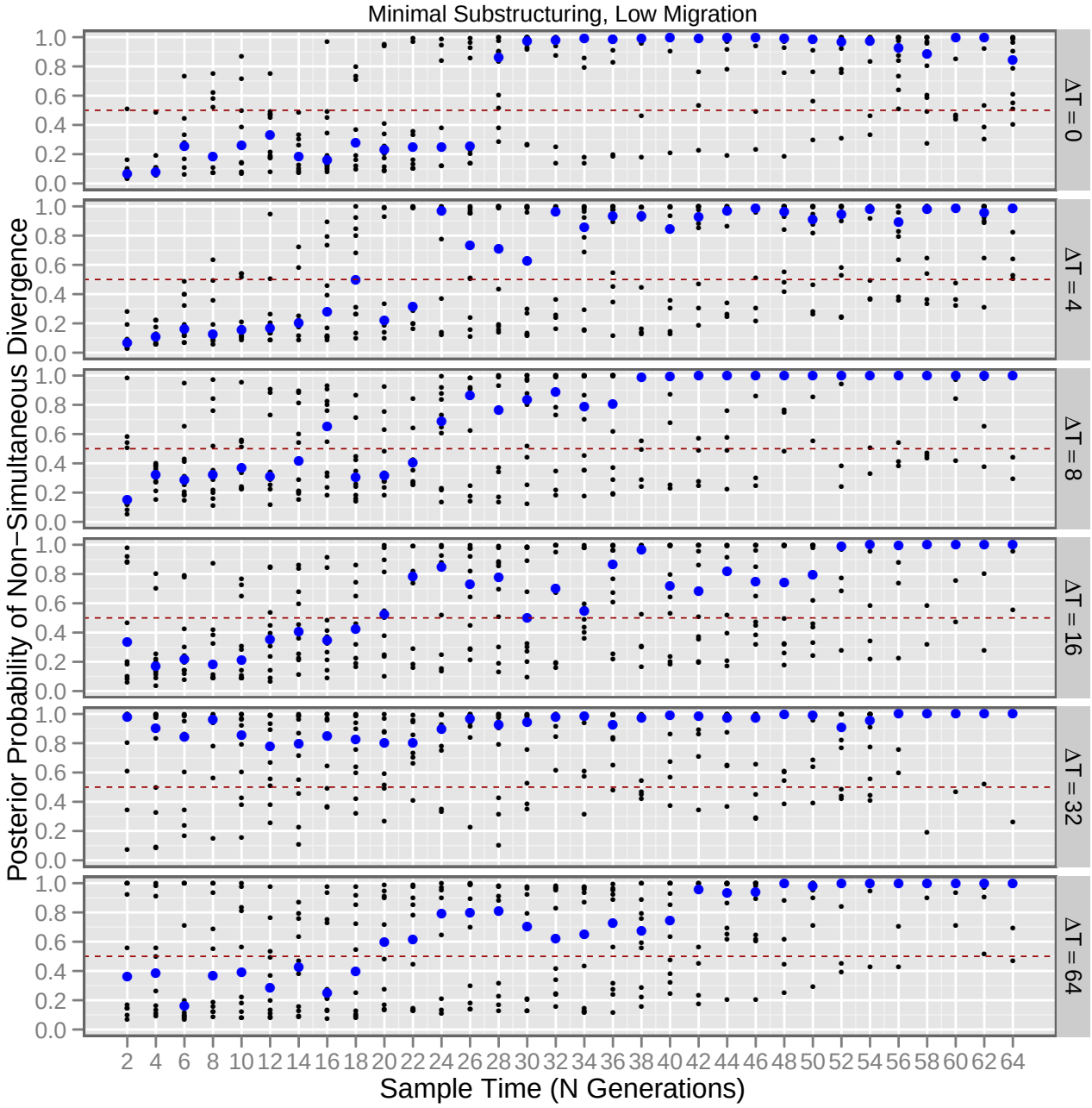


Figure 2.10: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *low* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

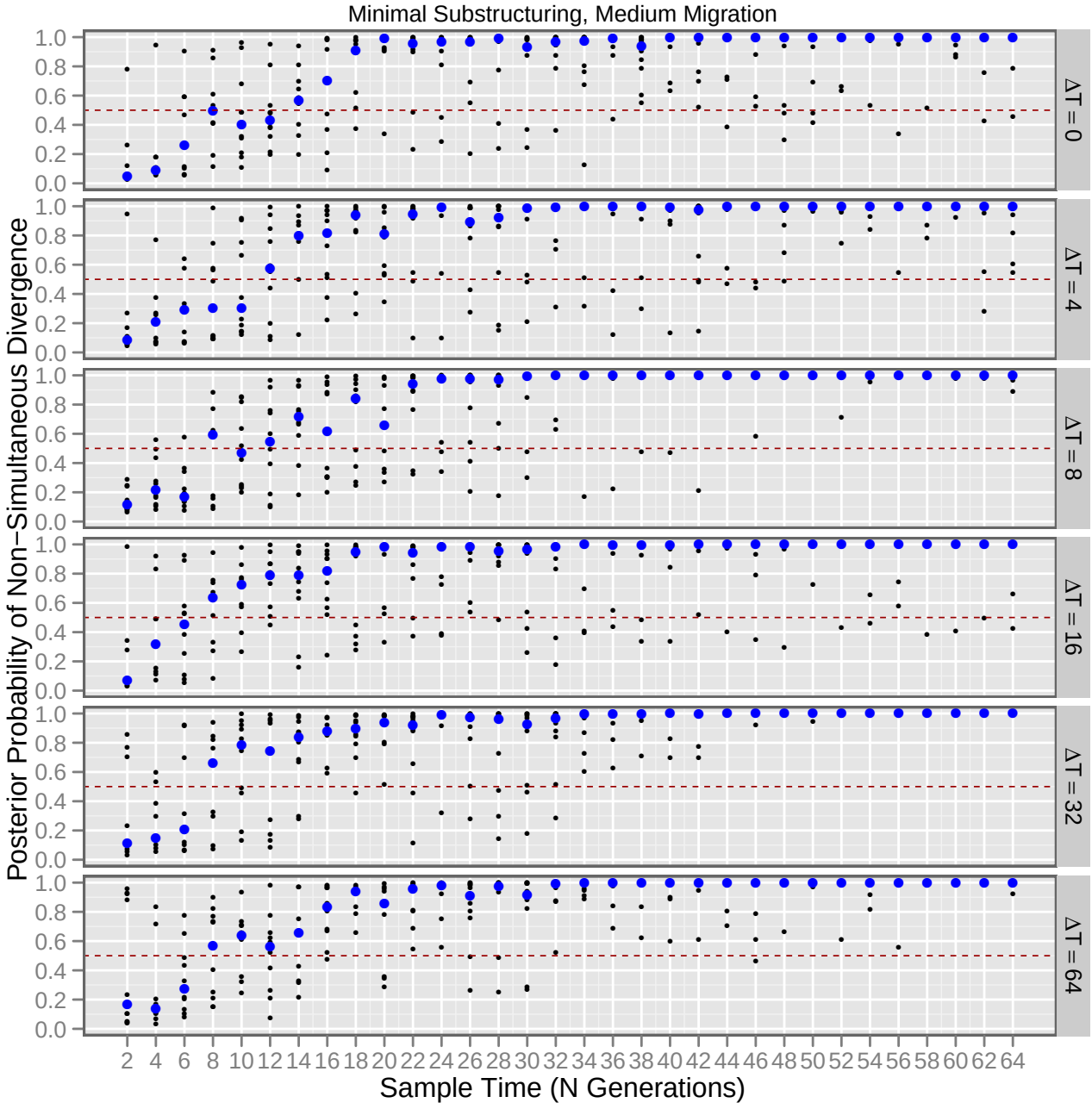


Figure 2.11: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *medium* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

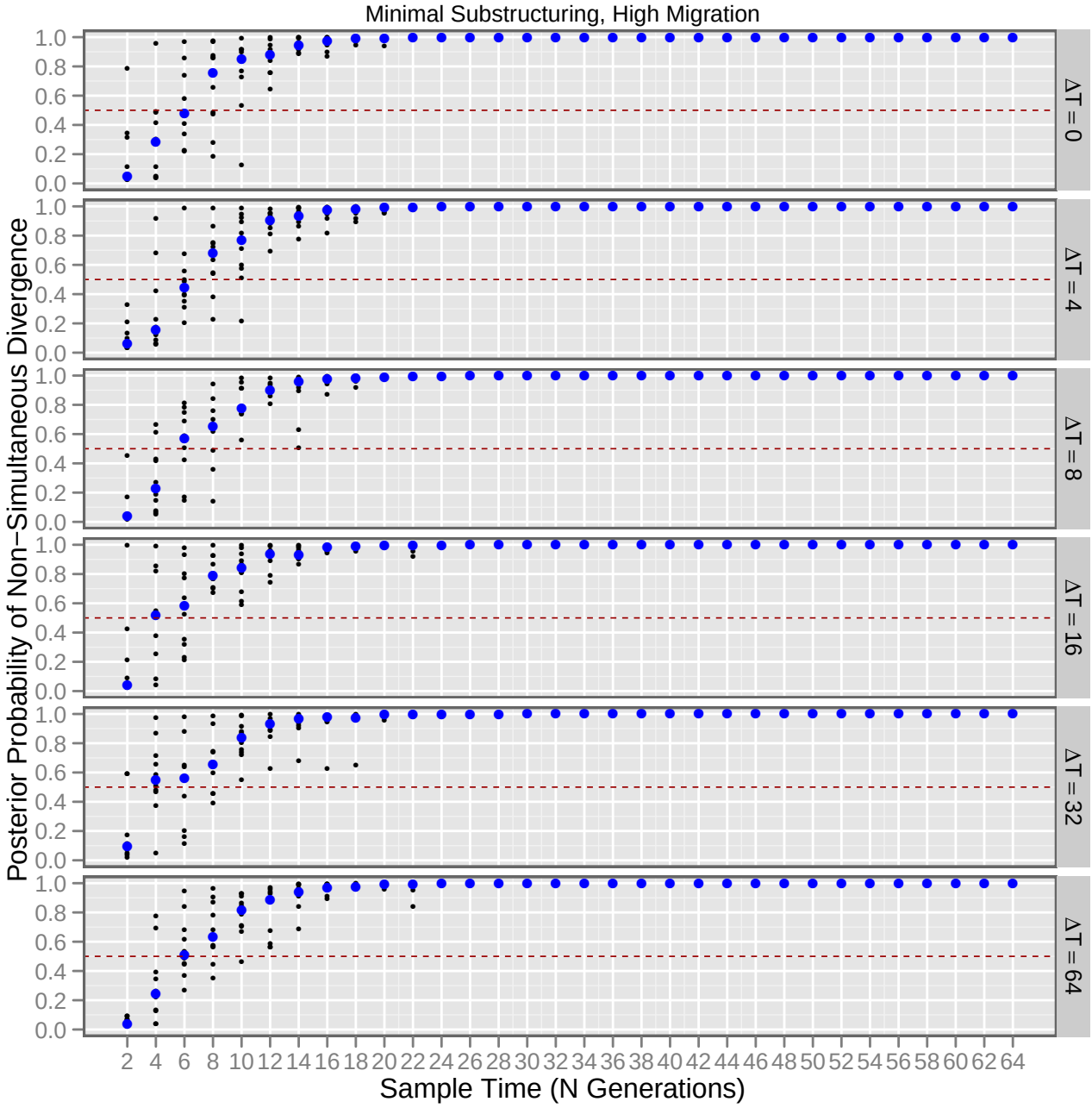


Figure 2.12: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *high* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

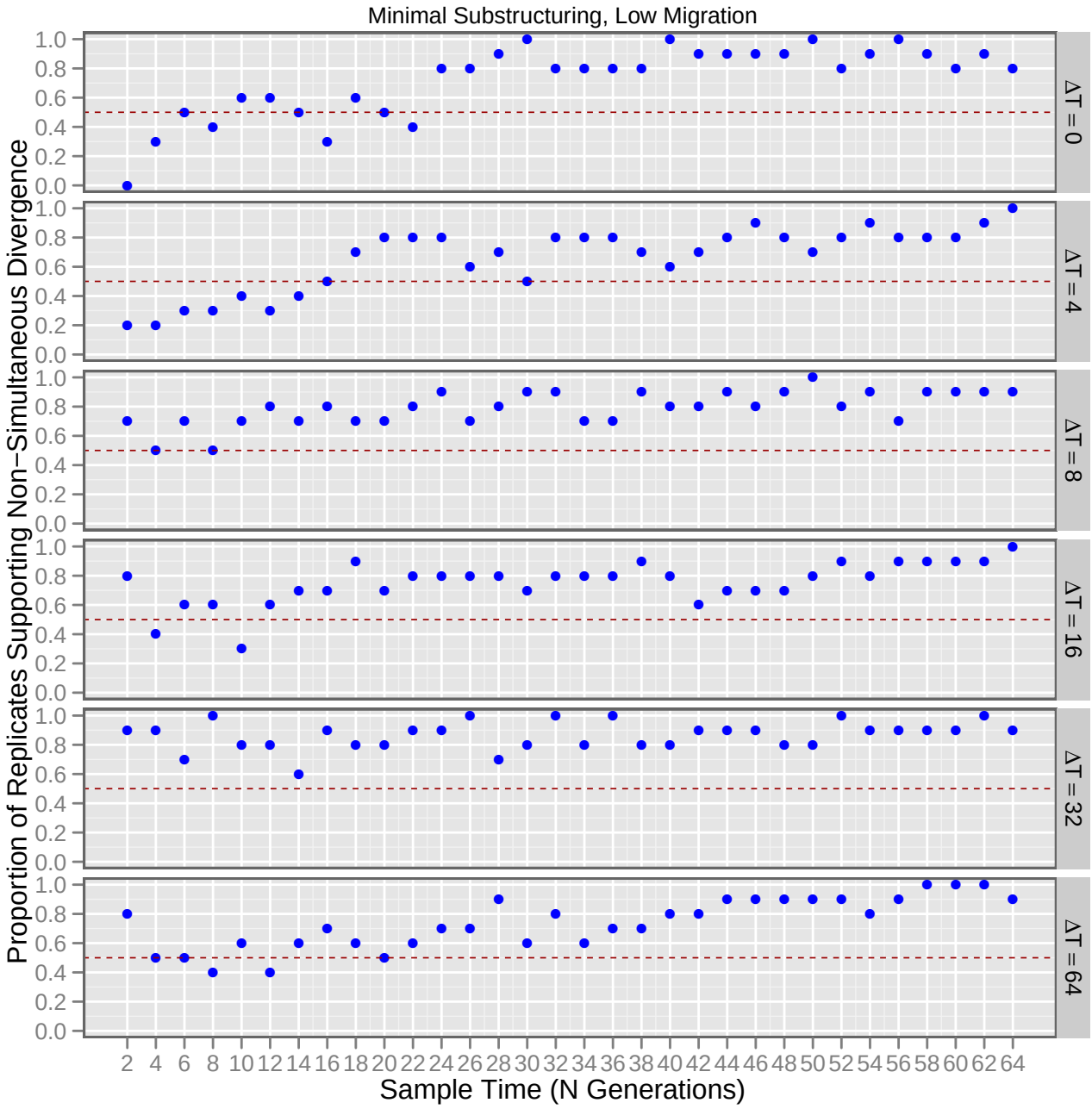


Figure 2.13: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *low post-vicariance migration* levels and *high* theta values when analyzed using a `msbayes` that does not account for migration, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

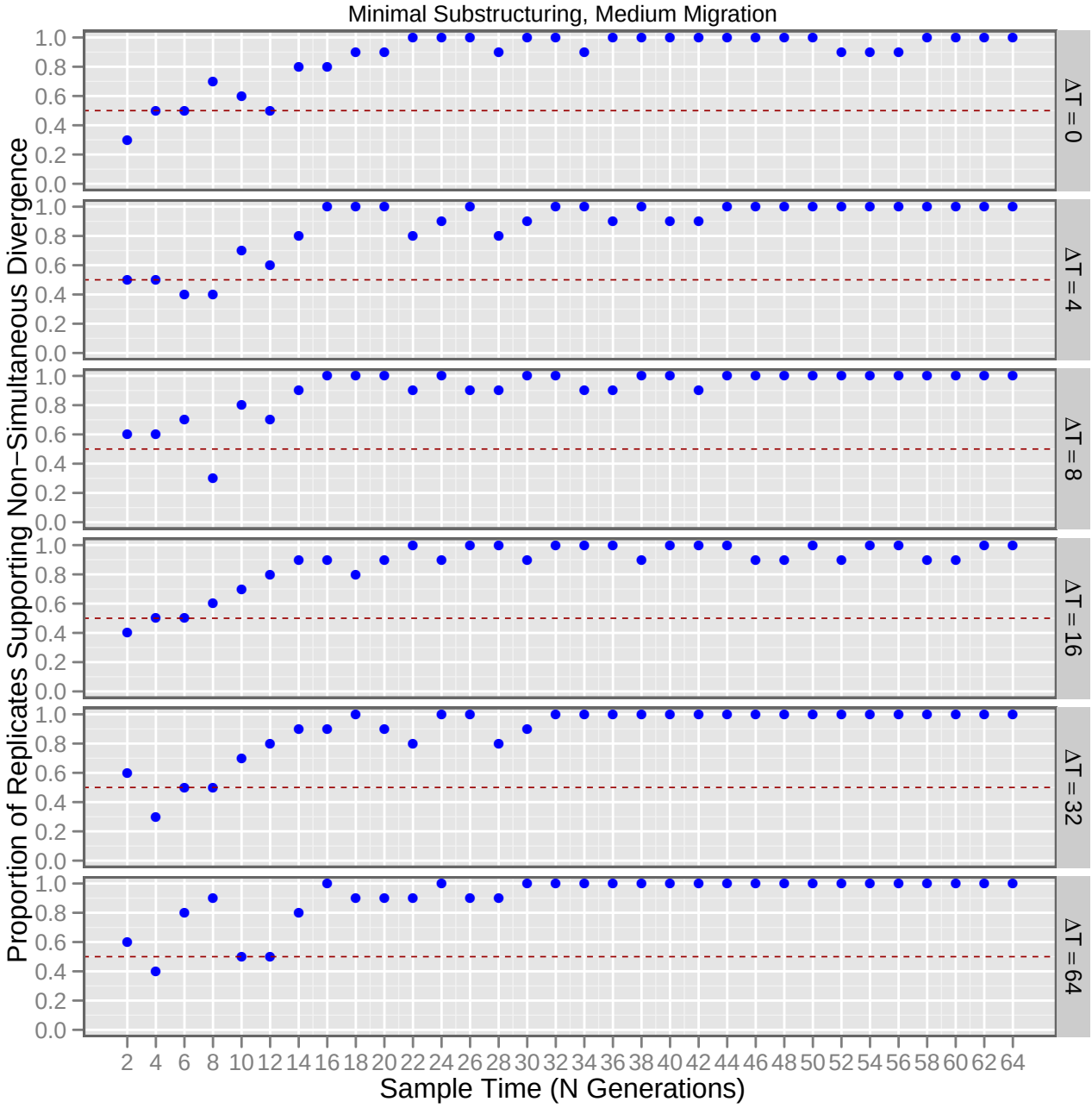


Figure 2.14: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *medium post-vicariance migration* levels and *high* theta values when analyzed using `msbayes` that does not account for migration, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

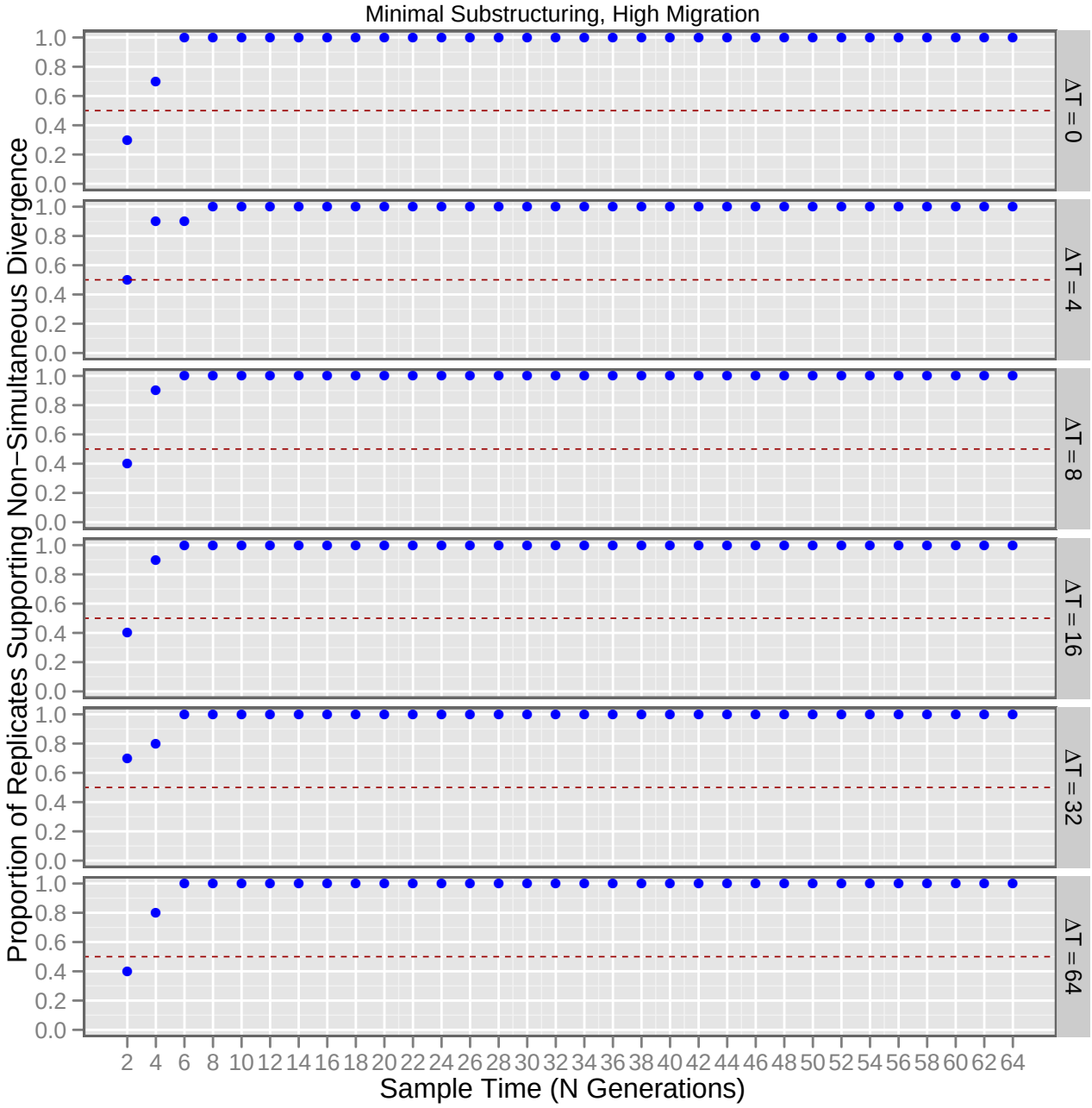


Figure 2.15: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *high post-vicariance migration* levels and *high* theta values when analyzed using `msbayes` that does not account for migration, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

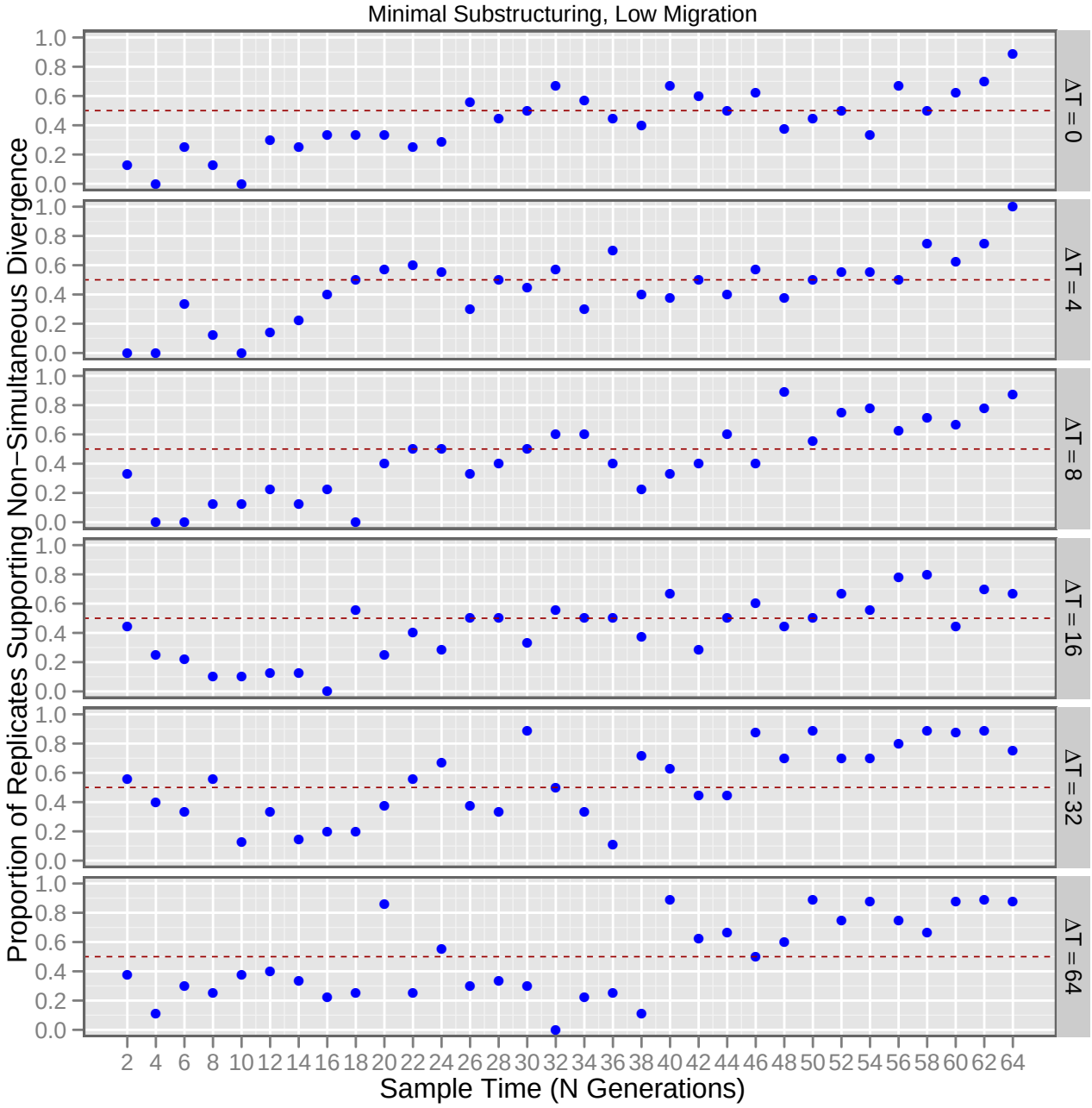


Figure 2.16: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *low post-vicariance migration* levels and *low* theta values when analyzed using a `msbayes` that does not account for migration, and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

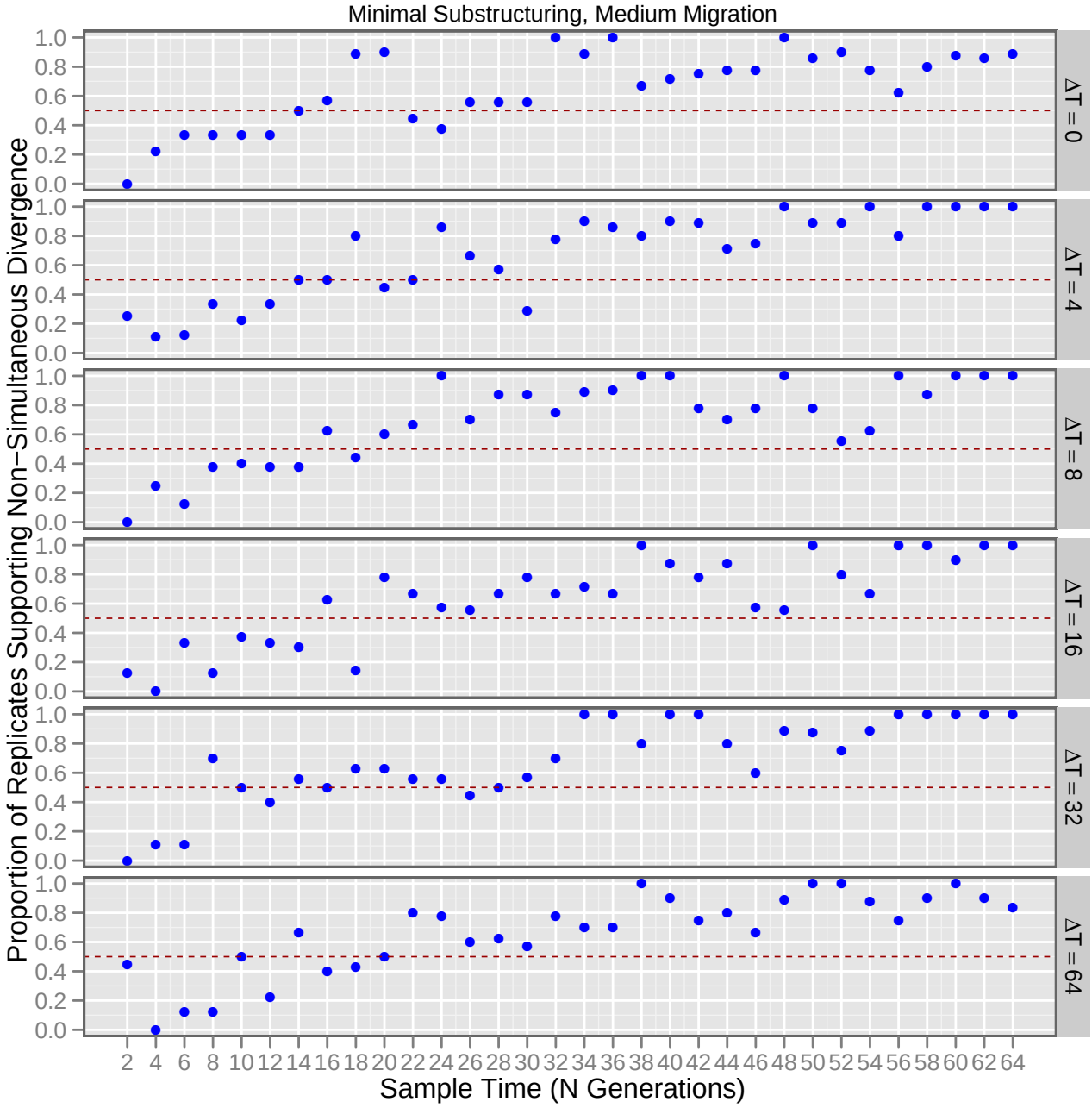


Figure 2.17: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *medium post-vicariance migration* levels and *low* theta values when analyzed using `msbayes` that does not account for migration, and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

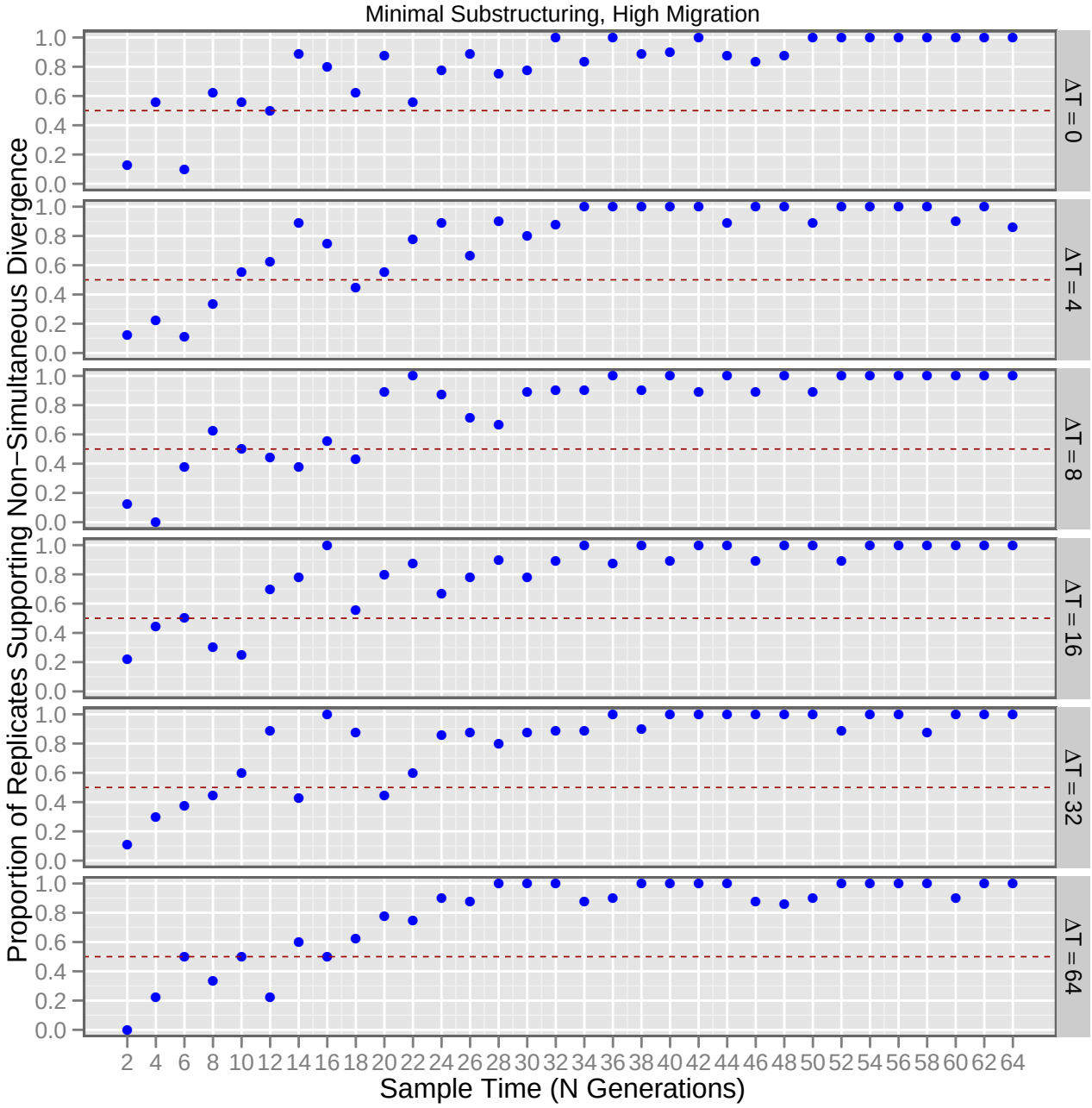


Figure 2.18: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *high post-vicariance migration* levels and *low* theta values when analyzed using `msbayes` that does not account for migration, and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

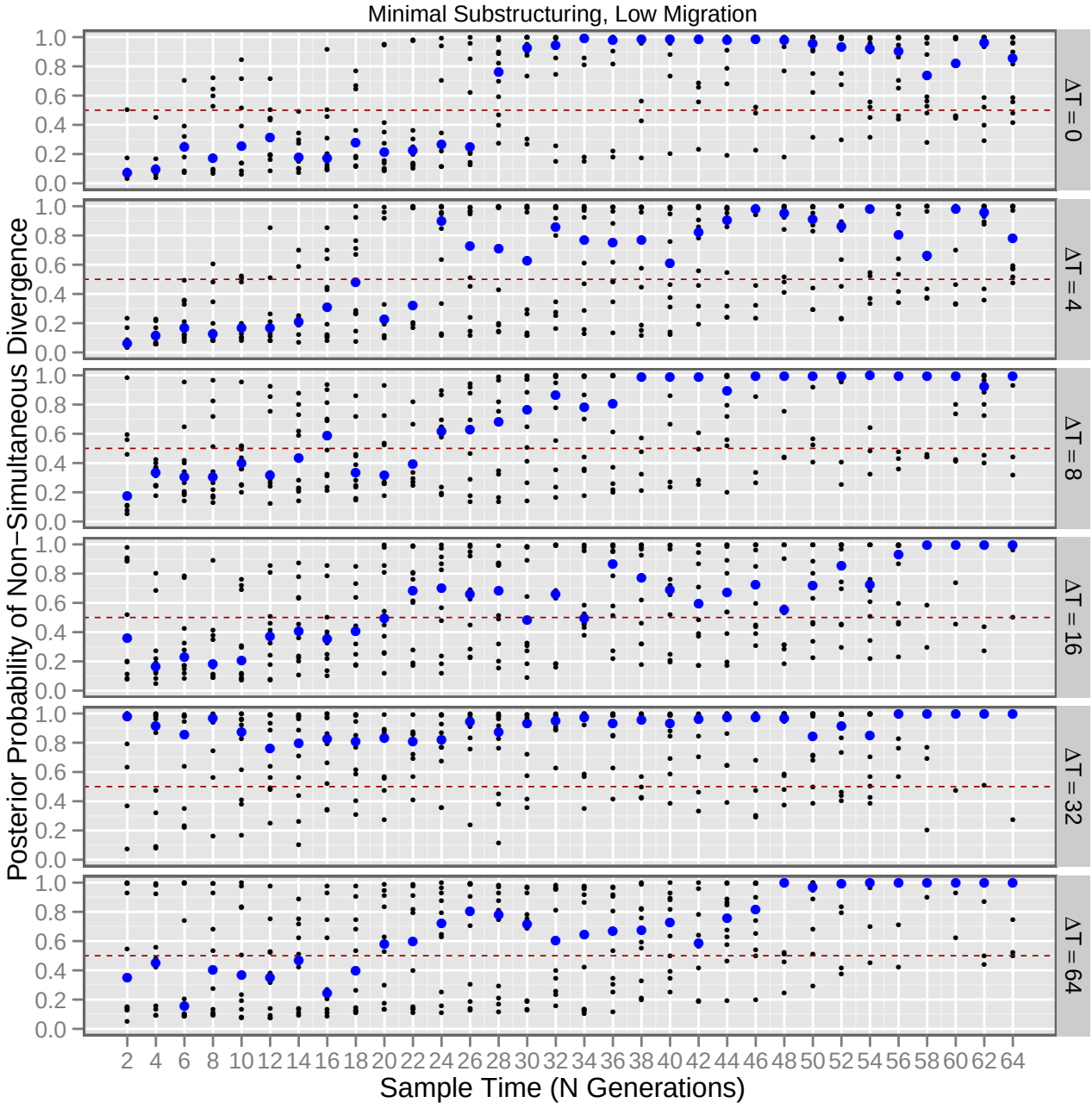


Figure 2.19: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *low* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that allows for *low* levels of migration. See figure 2.4 for details on interpreting the plots.

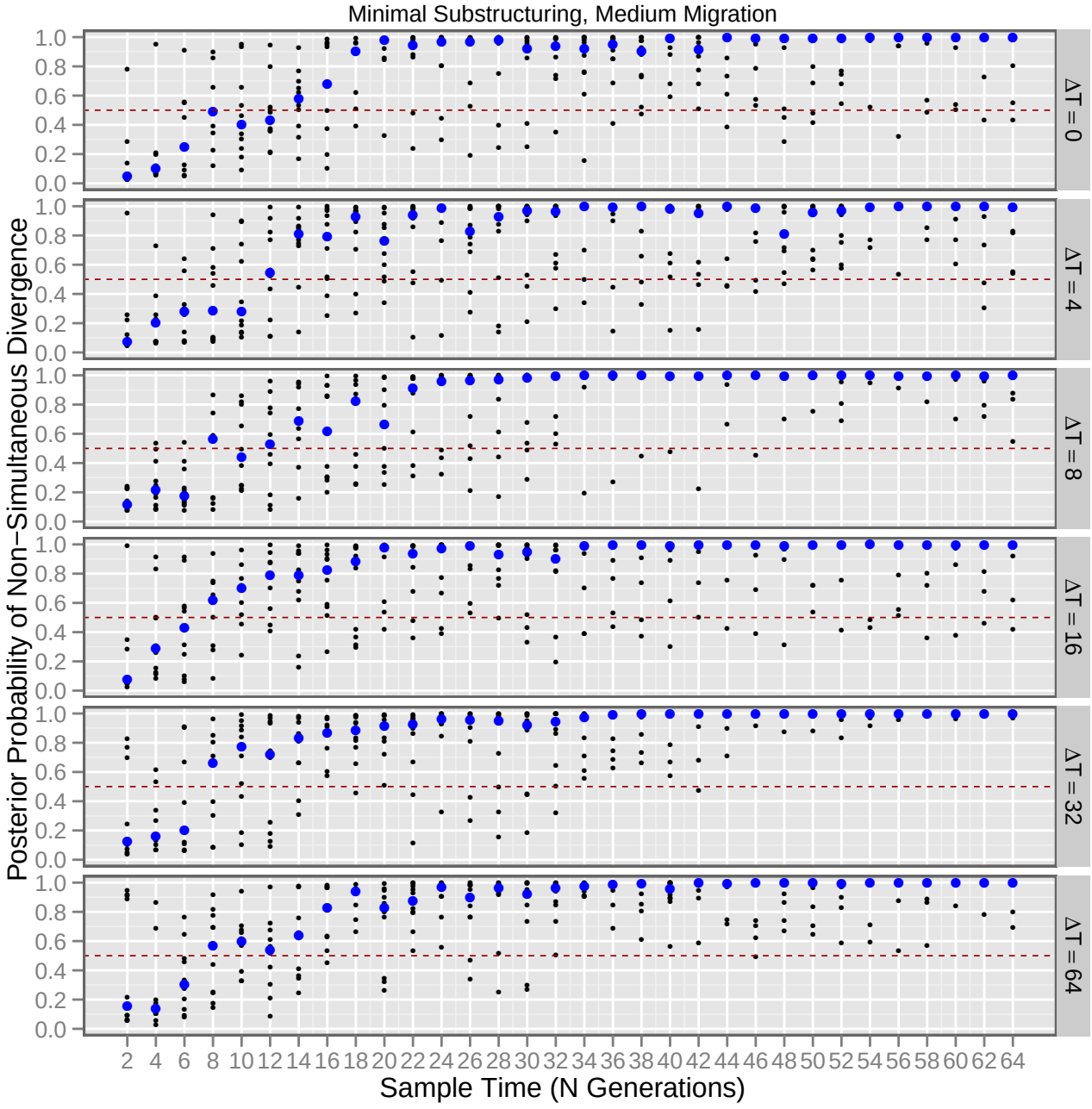


Figure 2.20: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *medium* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that allows for *low* levels of migration. See figure 2.4 for details on interpreting the plots.

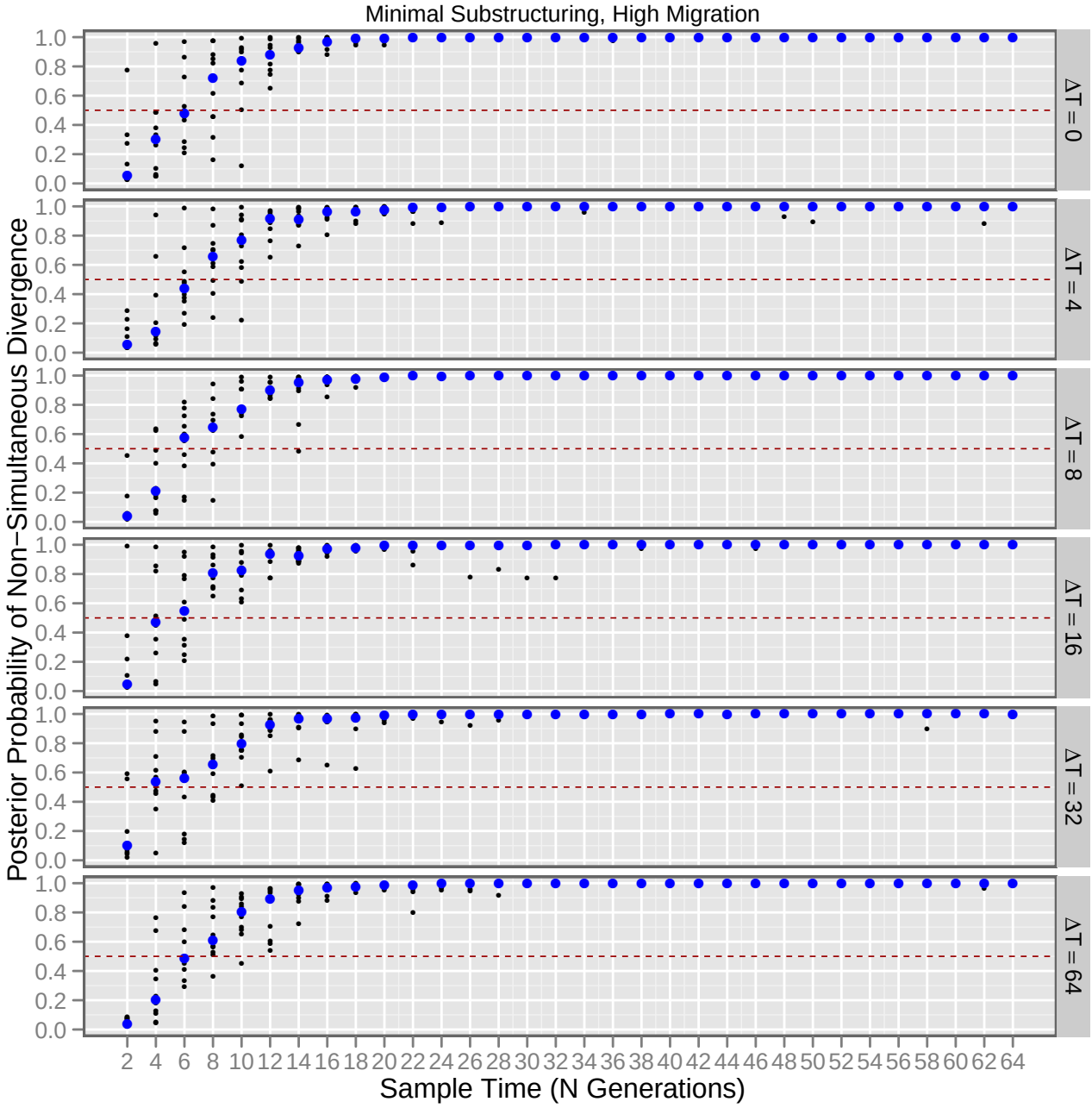


Figure 2.21: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *high* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that allows for *low* levels of migration. See figure 2.4 for details on interpreting the plots.

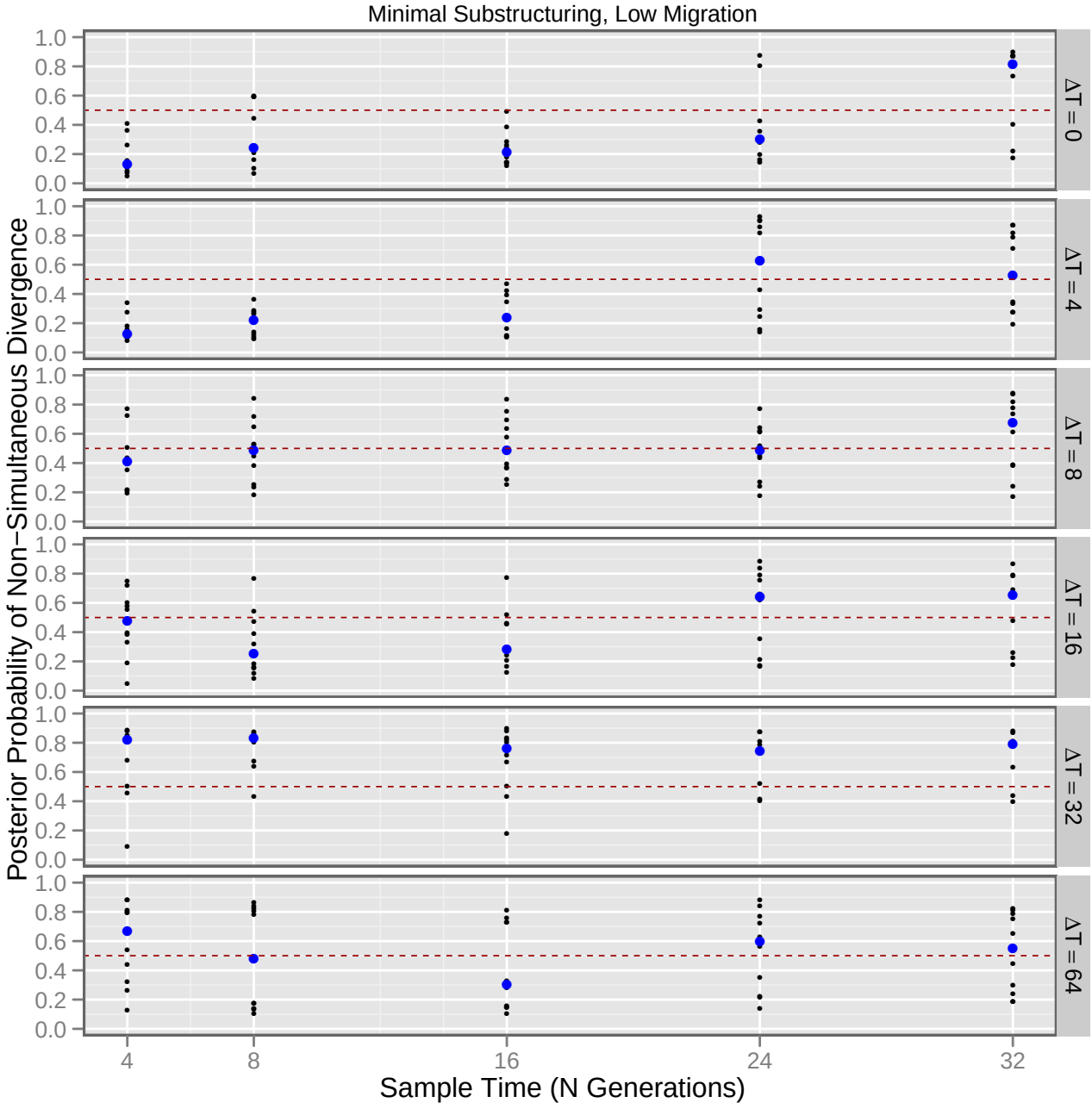


Figure 2.22: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *low* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that allows for *high* levels of migration. See figure 2.4 for details on interpreting the plots.

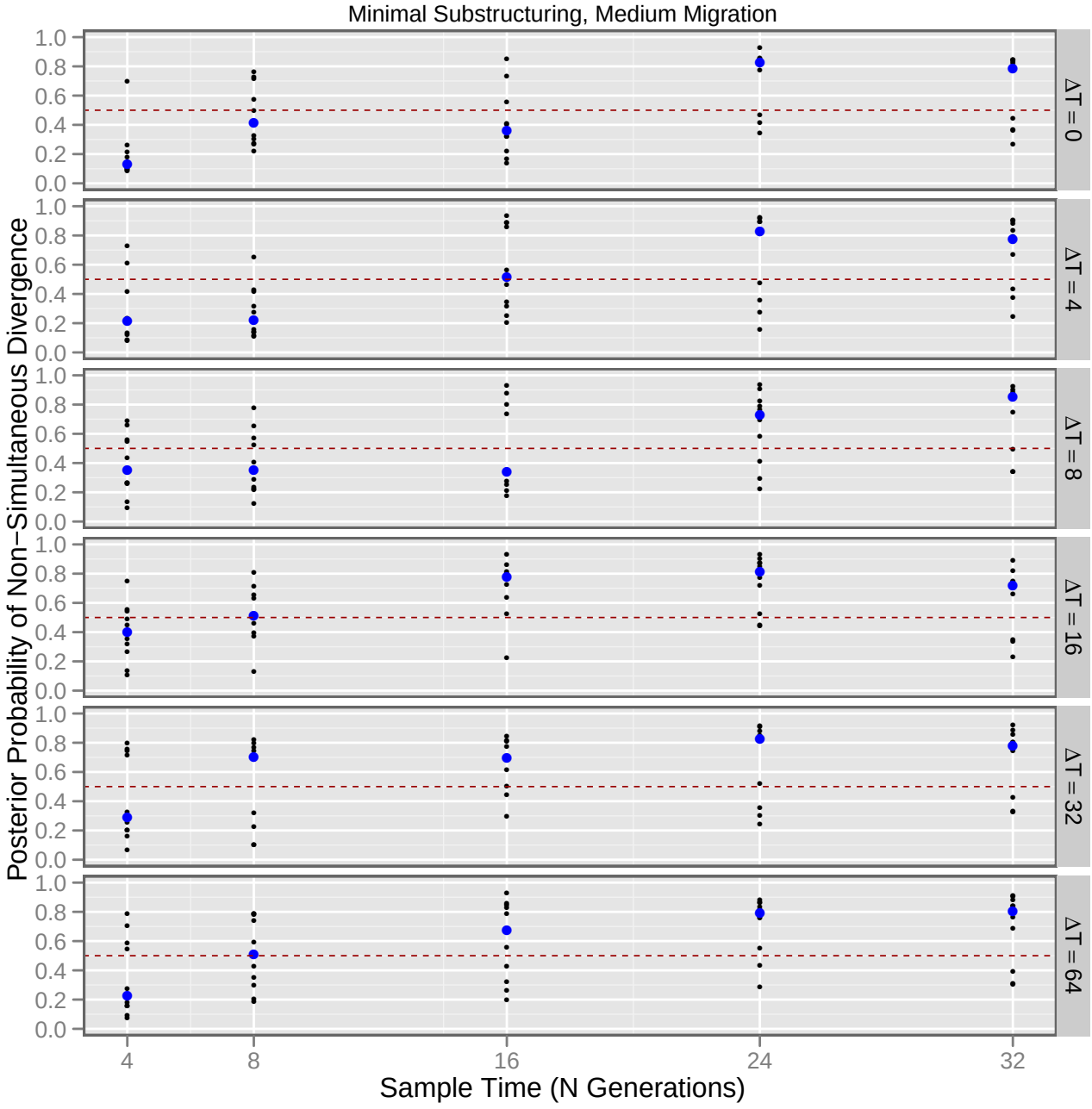


Figure 2.23: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *medium* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that allows for textithigh levels of migration. See figure 2.4 for details on interpreting the plots.

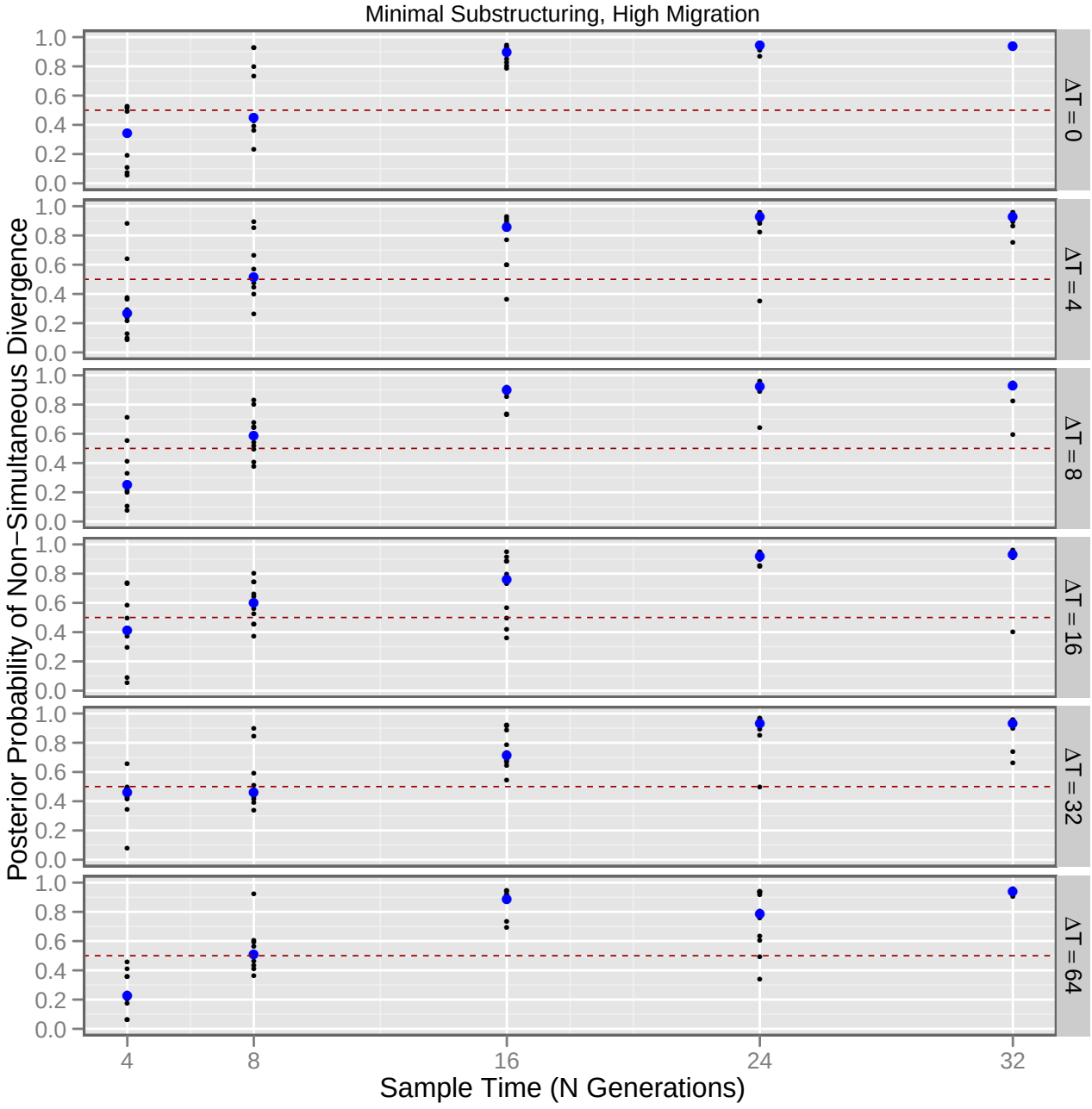


Figure 2.24: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *high* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that allows for textithigh levels of migration. See figure 2.4 for details on interpreting the plots.

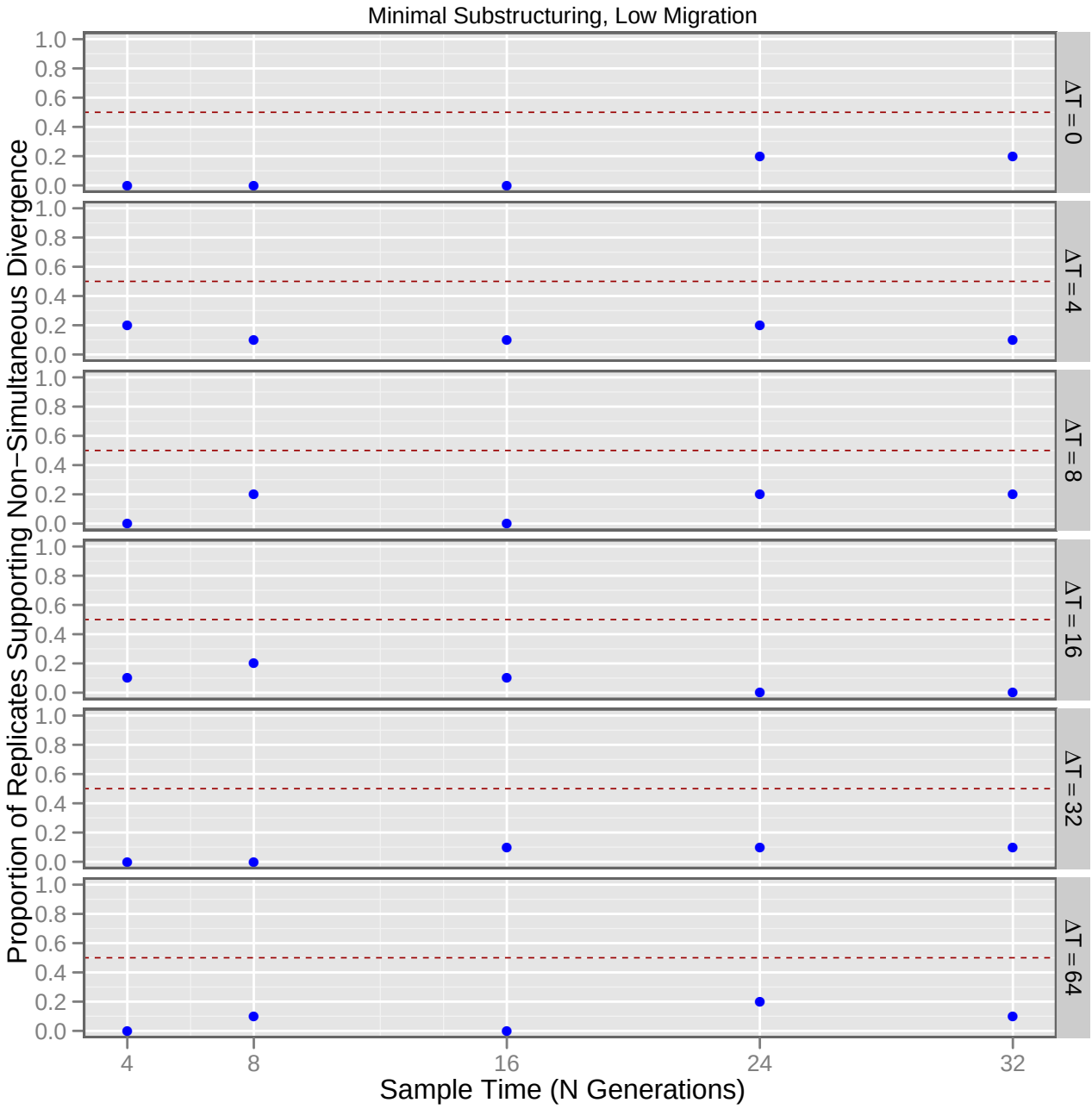


Figure 2.25: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *low post-vicariance migration* levels and *high* theta values when analyzed using a `msbayes` that assumes high levels of migration, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

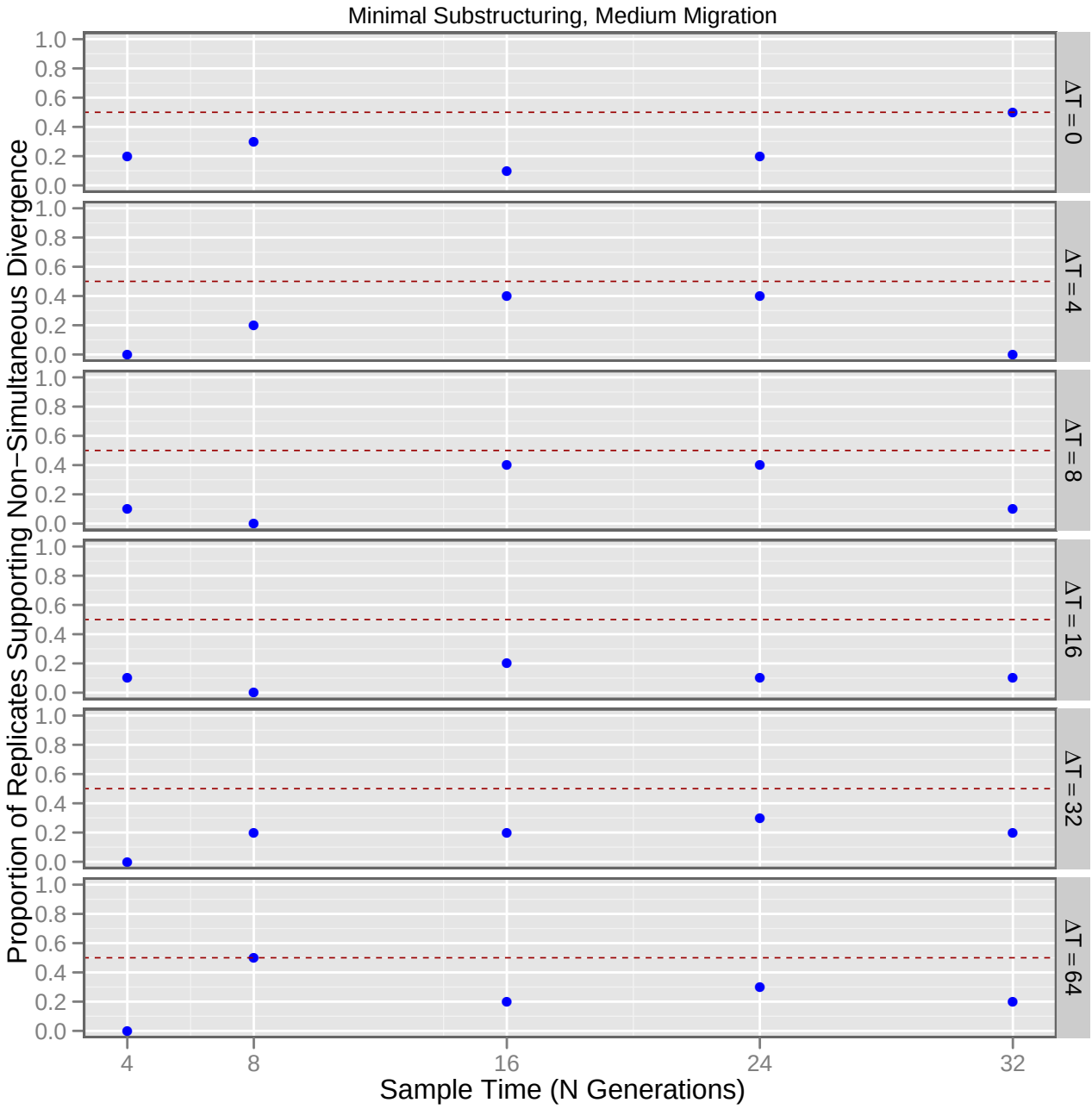


Figure 2.26: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *medium post-vicariance migration* levels and *high* theta values when analyzed using `msbayes` that assumes high levels of migration, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

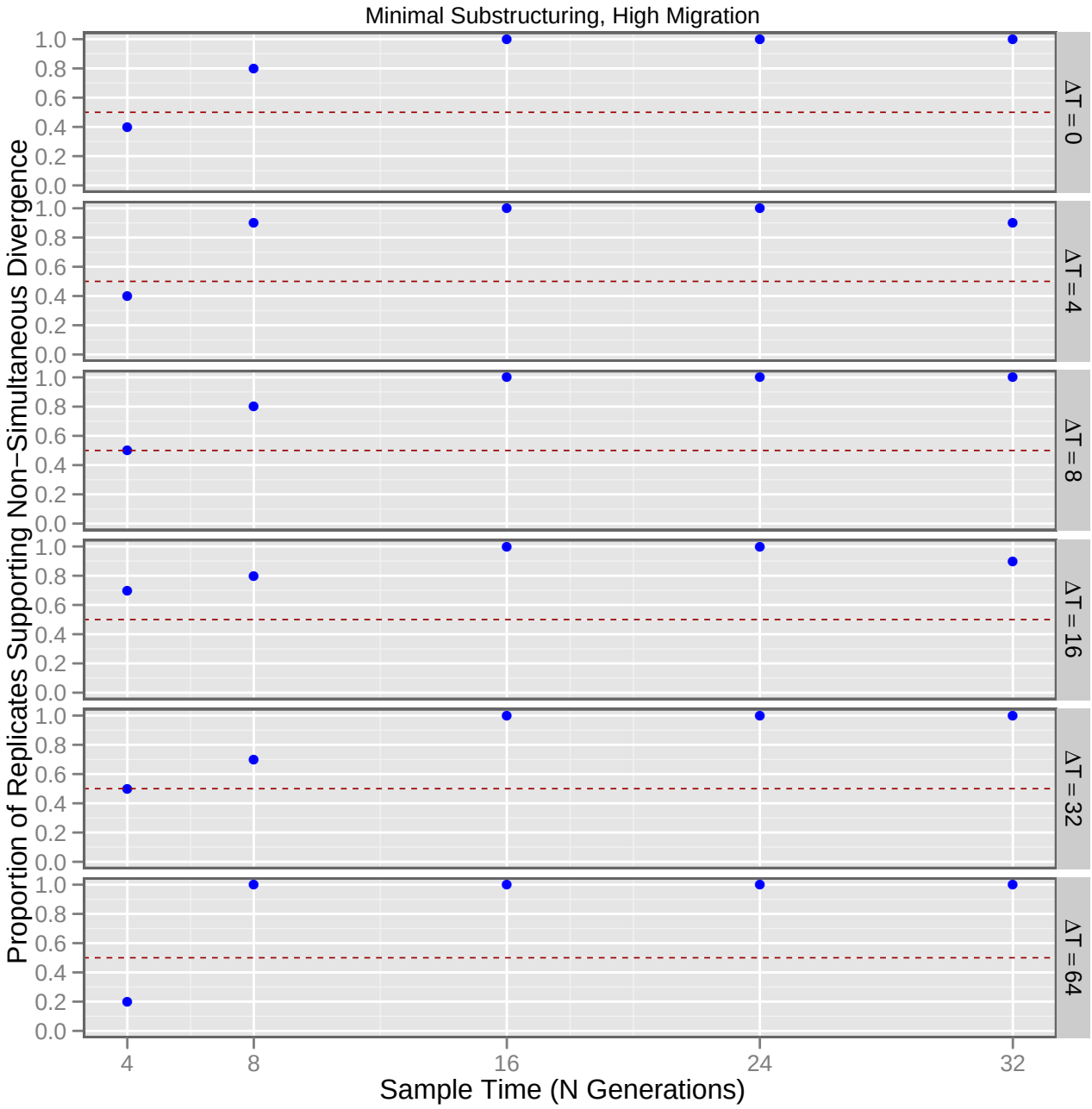


Figure 2.27: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *high post-vicariance migration* levels and *high* theta values when analyzed using `msbayes` that assumes high levels of migration, and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

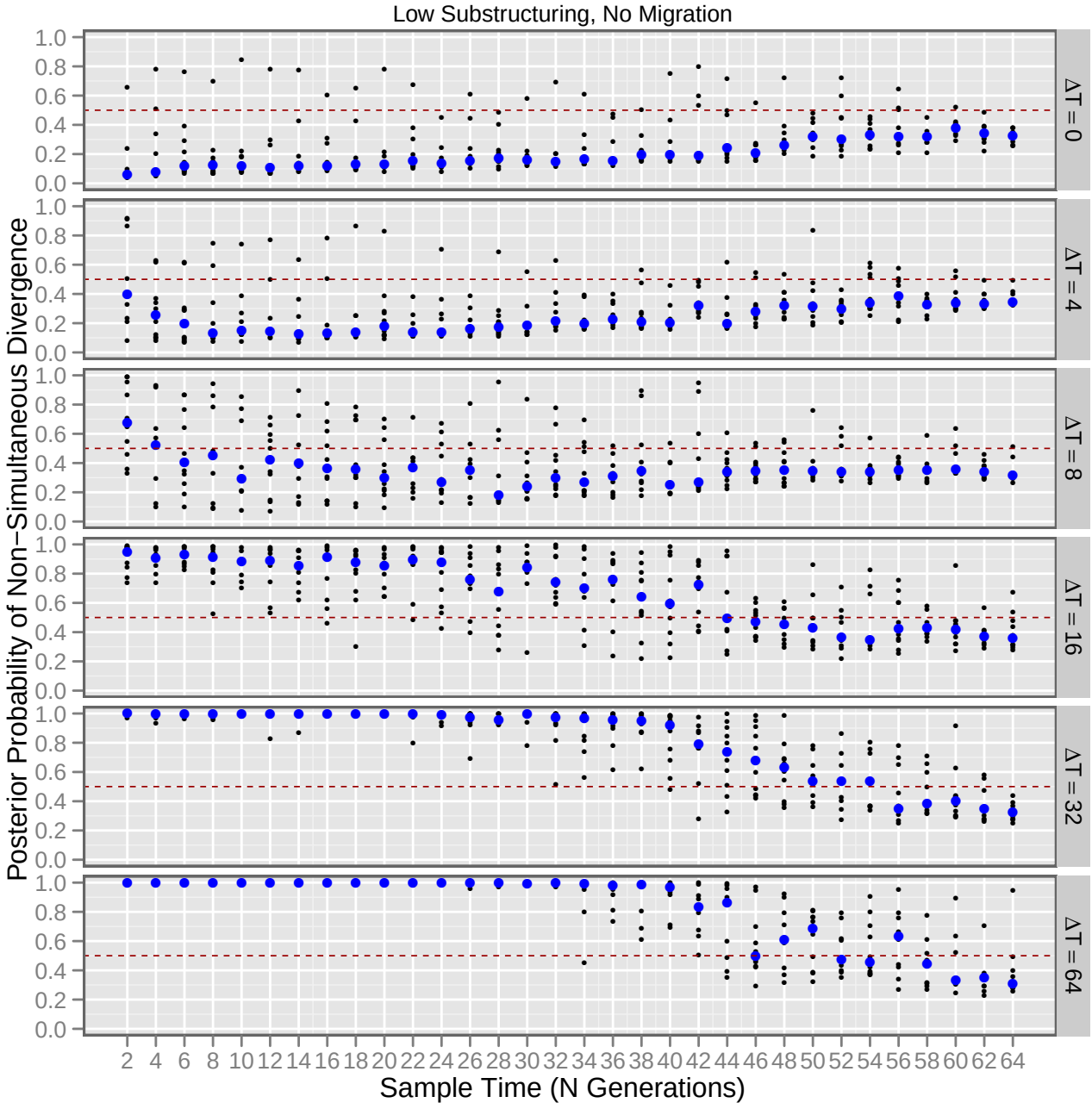


Figure 2.28: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *low* levels of within population structuring when analyzed using `msbayes`. See figure 2.4 for details on interpreting the plots.

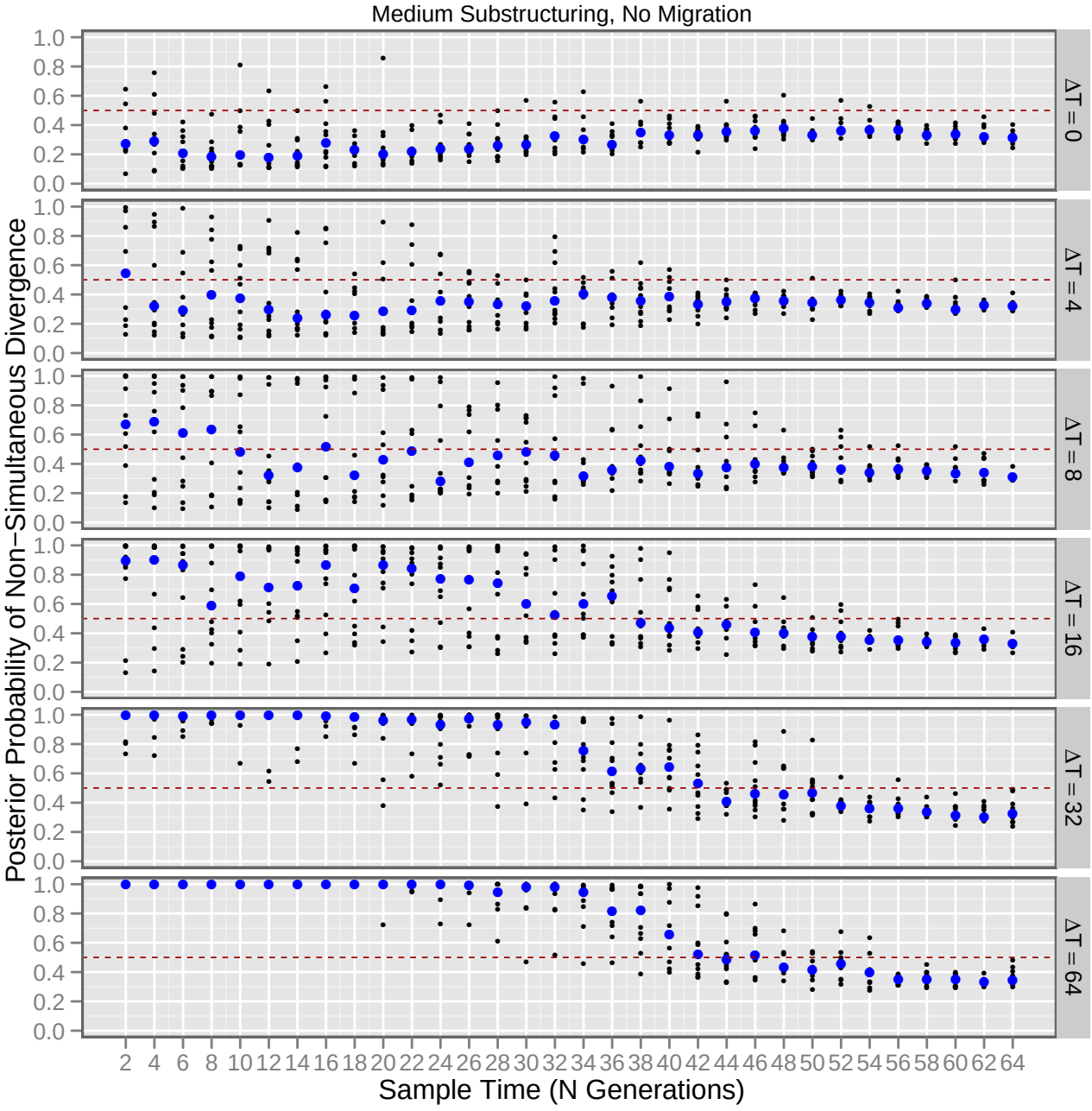


Figure 2.29: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *medium* levels of within population structuring when analyzed using `msbayes`. See figure 2.4 for details on interpreting the plots.

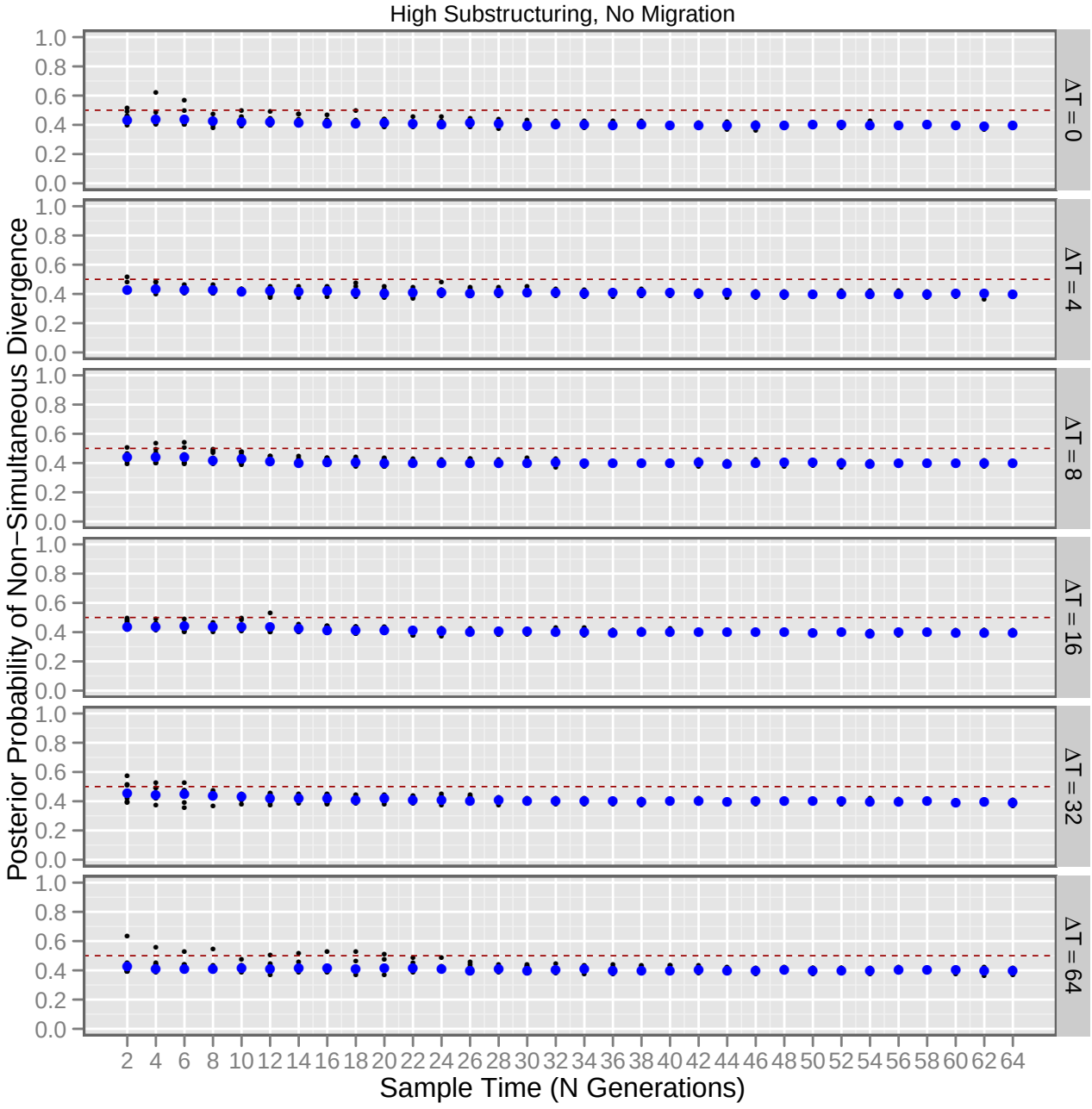


Figure 2.30: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values with *high* levels of within population structuring when analyzed using `msbayes`. See figure 2.4 for details on interpreting the plots.

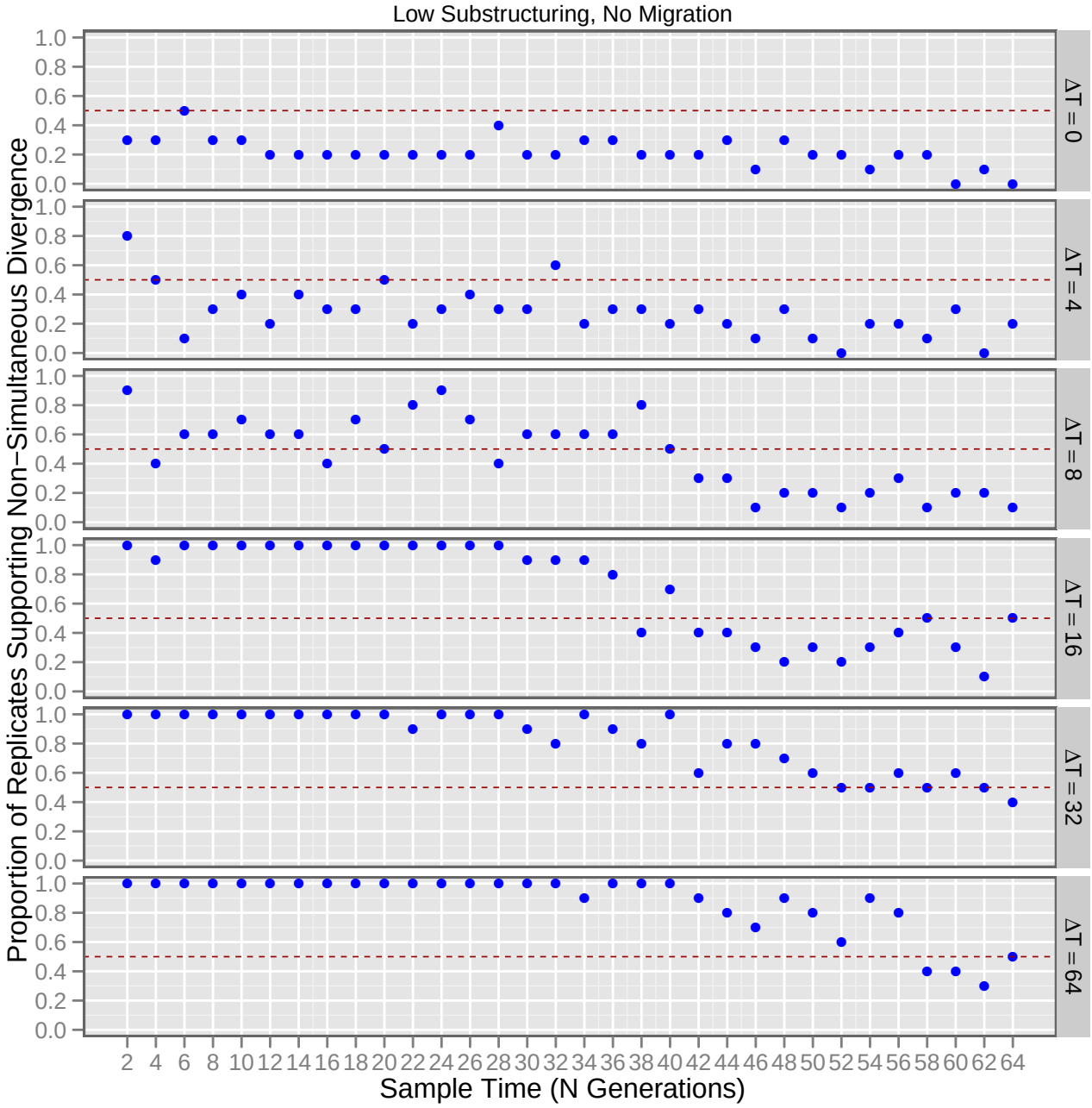


Figure 2.31: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *high* theta values and *low* levels *within-population substructuring* when analyzed using `msbayes` and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

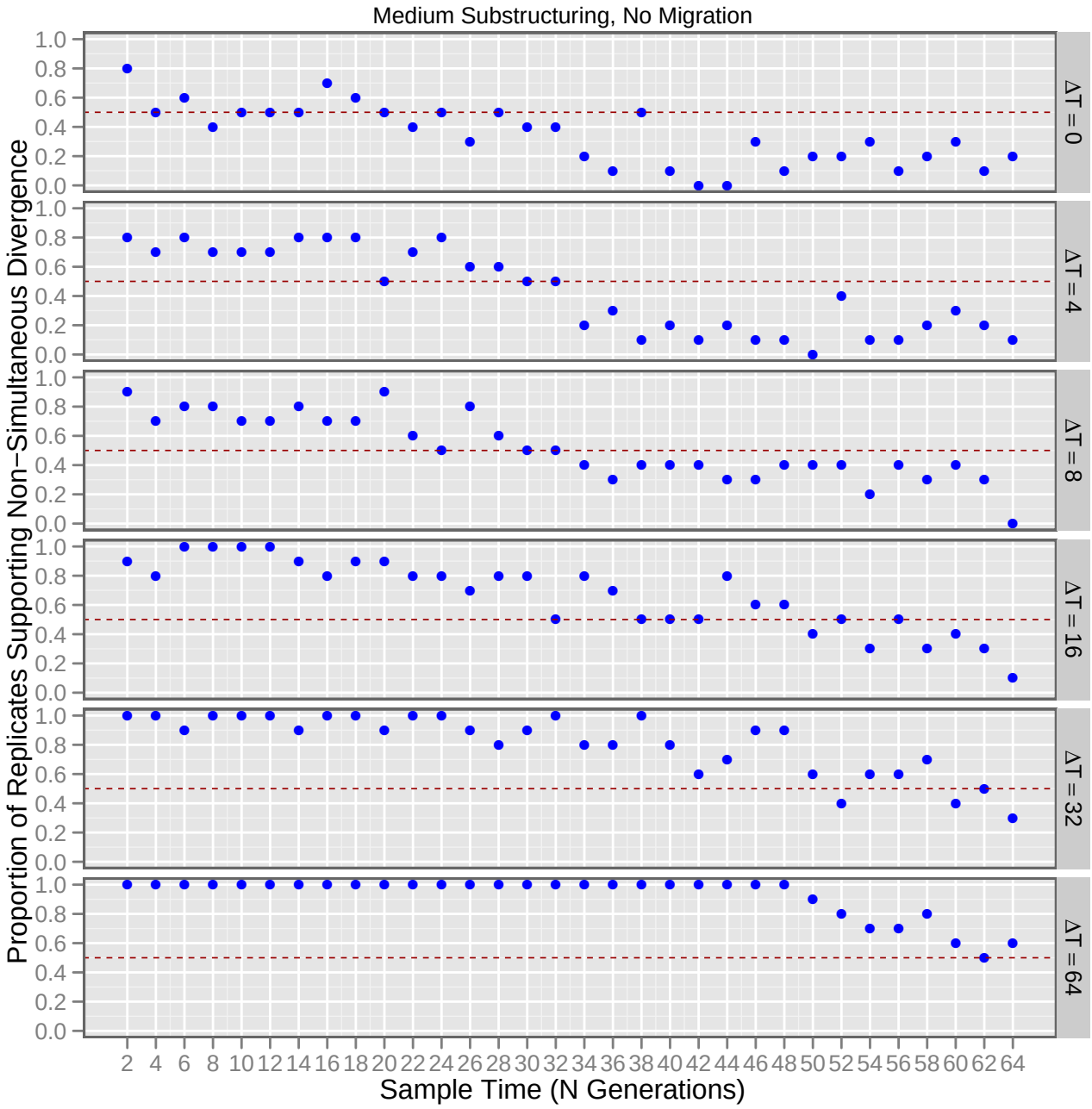


Figure 2.32: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *high* theta values and *medium* levels *within-population substructuring* when analyzed using `msbayes` and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

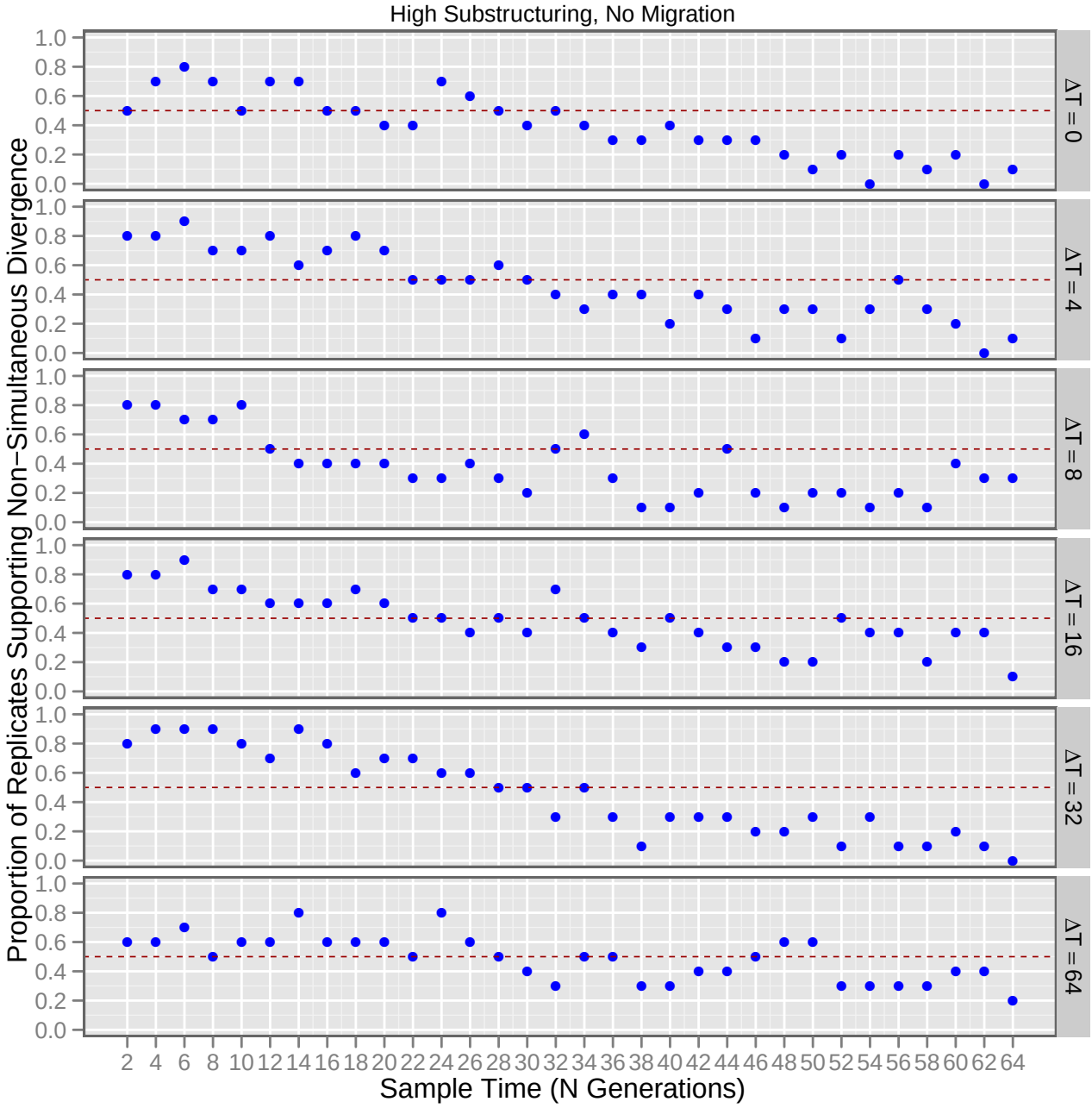


Figure 2.33: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *high* theta values and *high* levels *within-population substructuring* when analyzed using `msbayes` and using *weighted mode of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time model. See figure 2.4 for details on interpreting the plots.

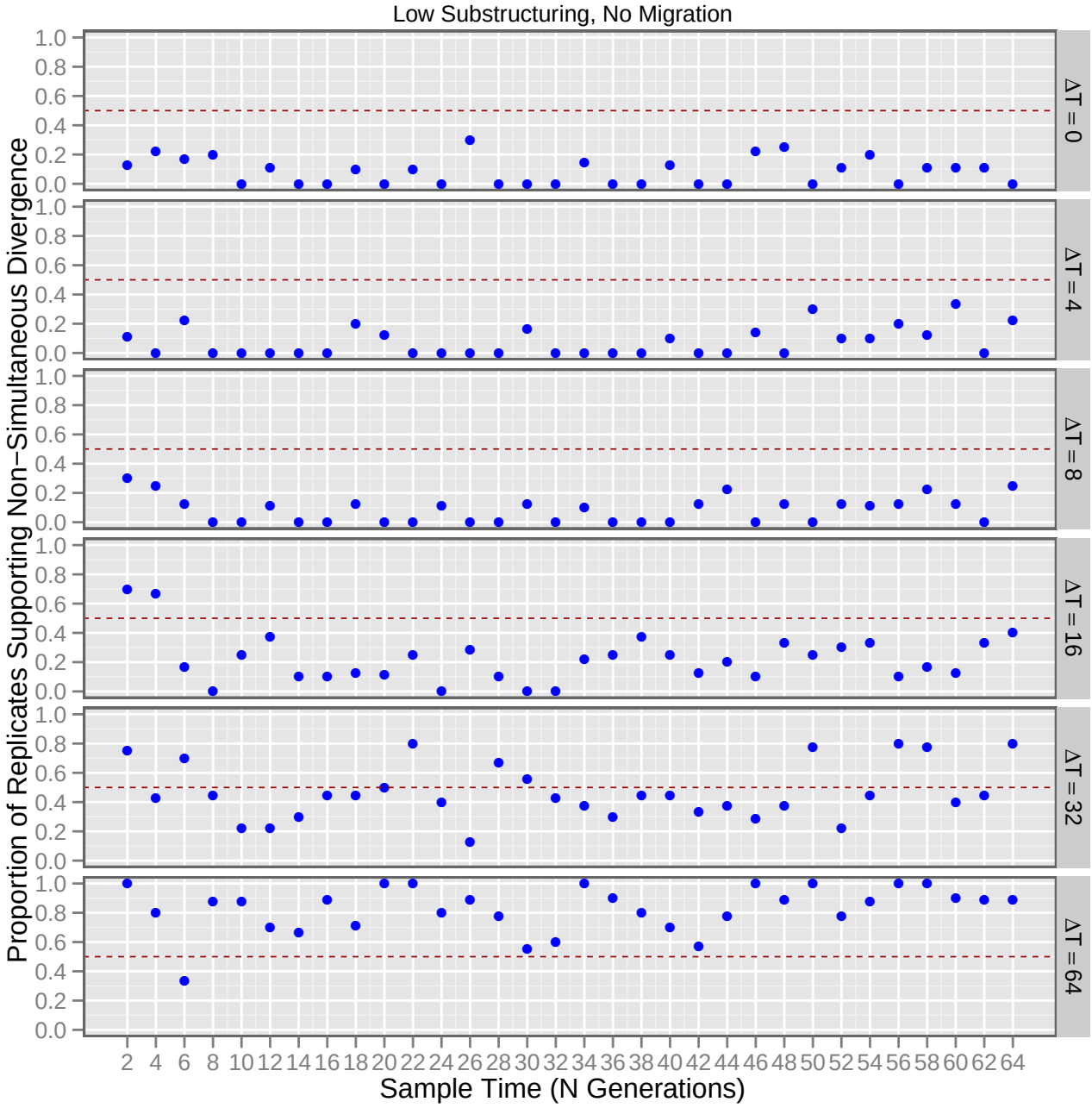


Figure 2.34: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *low* theta values and *low* levels *within-population substructuring* when analyzed using `msbayes` and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time meanl. See figure 2.4 for details on interpreting the plots.

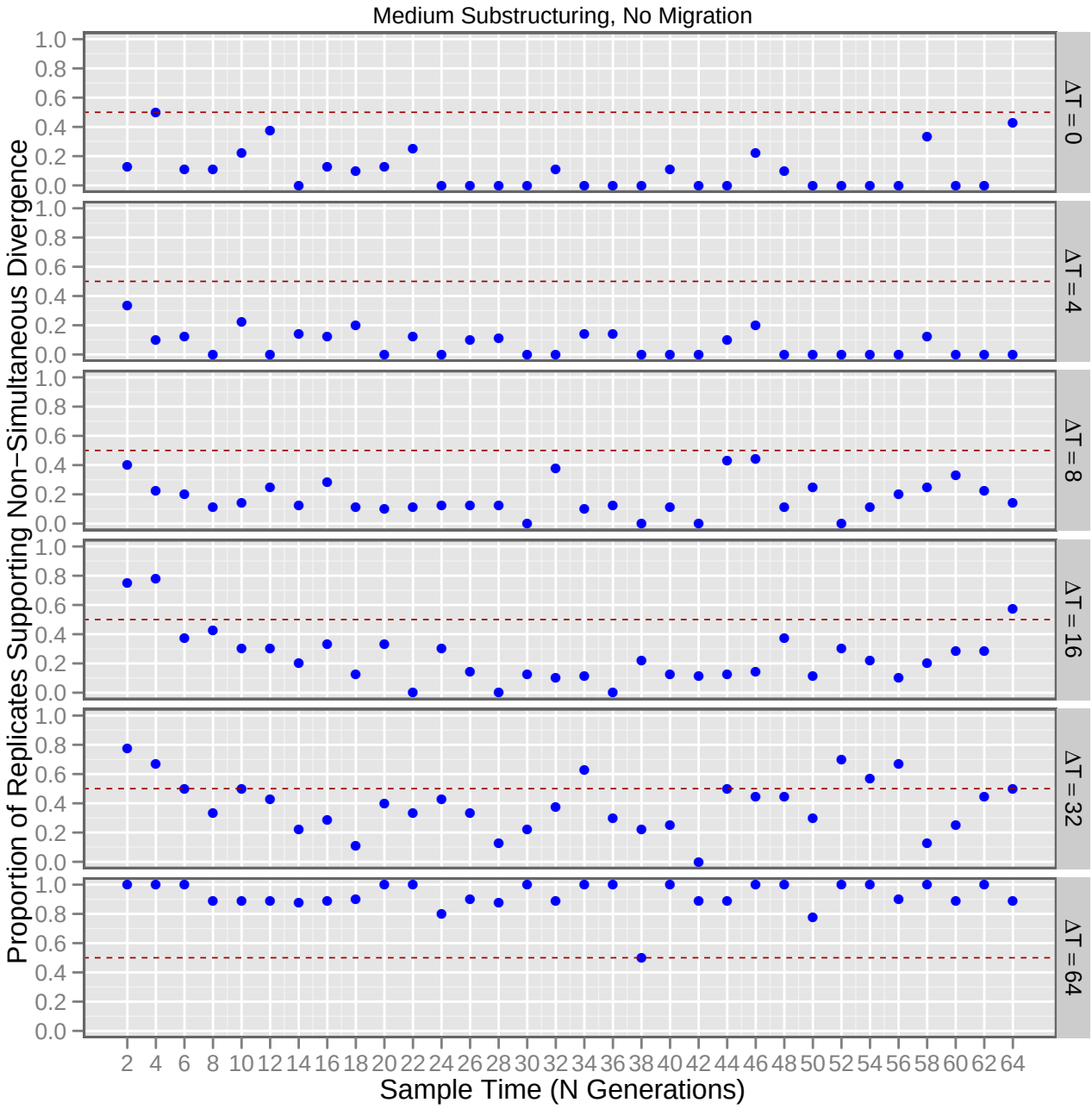


Figure 2.35: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *low* theta values and *medium* levels *within-population substructuring* when analyzed using `msbayes` and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time meanl. See figure 2.4 for details on interpreting the plots.

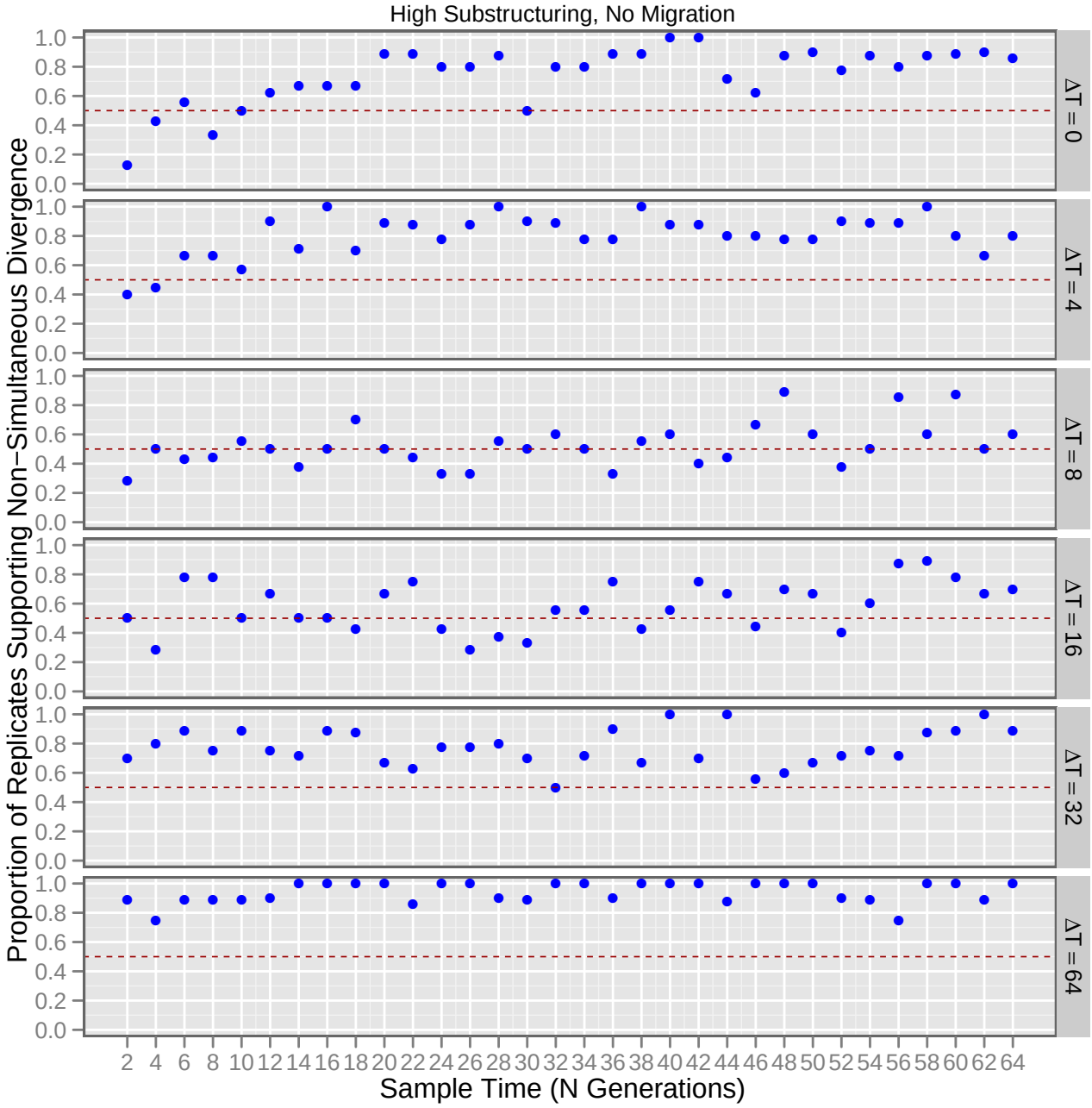


Figure 2.36: Proportion of replicates supporting non-simultaneous divergence of forward-time simulations under *low* theta values and *high* levels *within-population substructuring* when analyzed using `msbayes` and using *weighted mean of* Ω (the variance in divergence times divided by the mean; see text for details) to determine the divergence time meanl. See figure 2.4 for details on interpreting the plots.

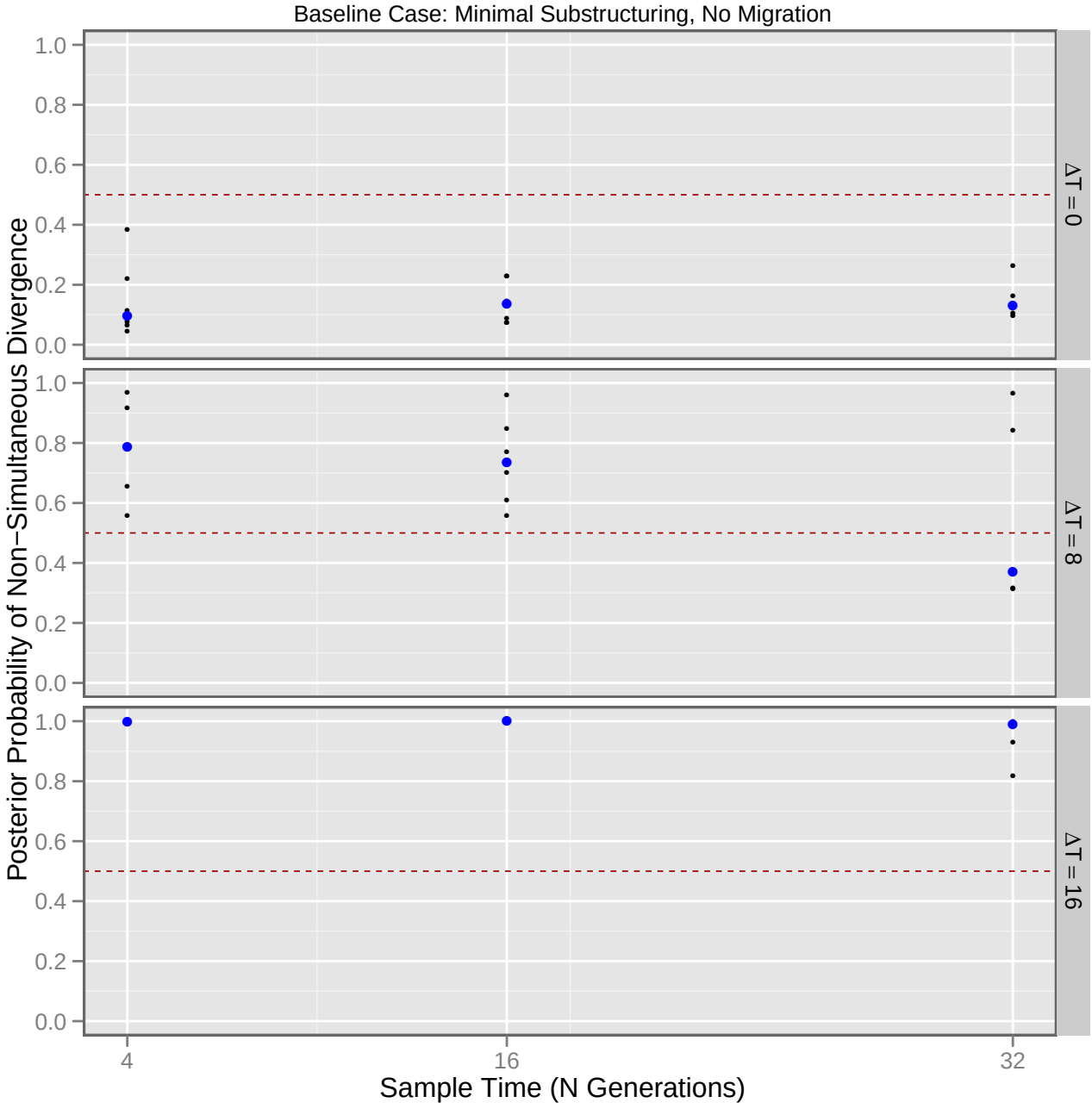


Figure 2.37: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under high theta values *for 5 independent loci* when analyzed using `msbayes`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met.

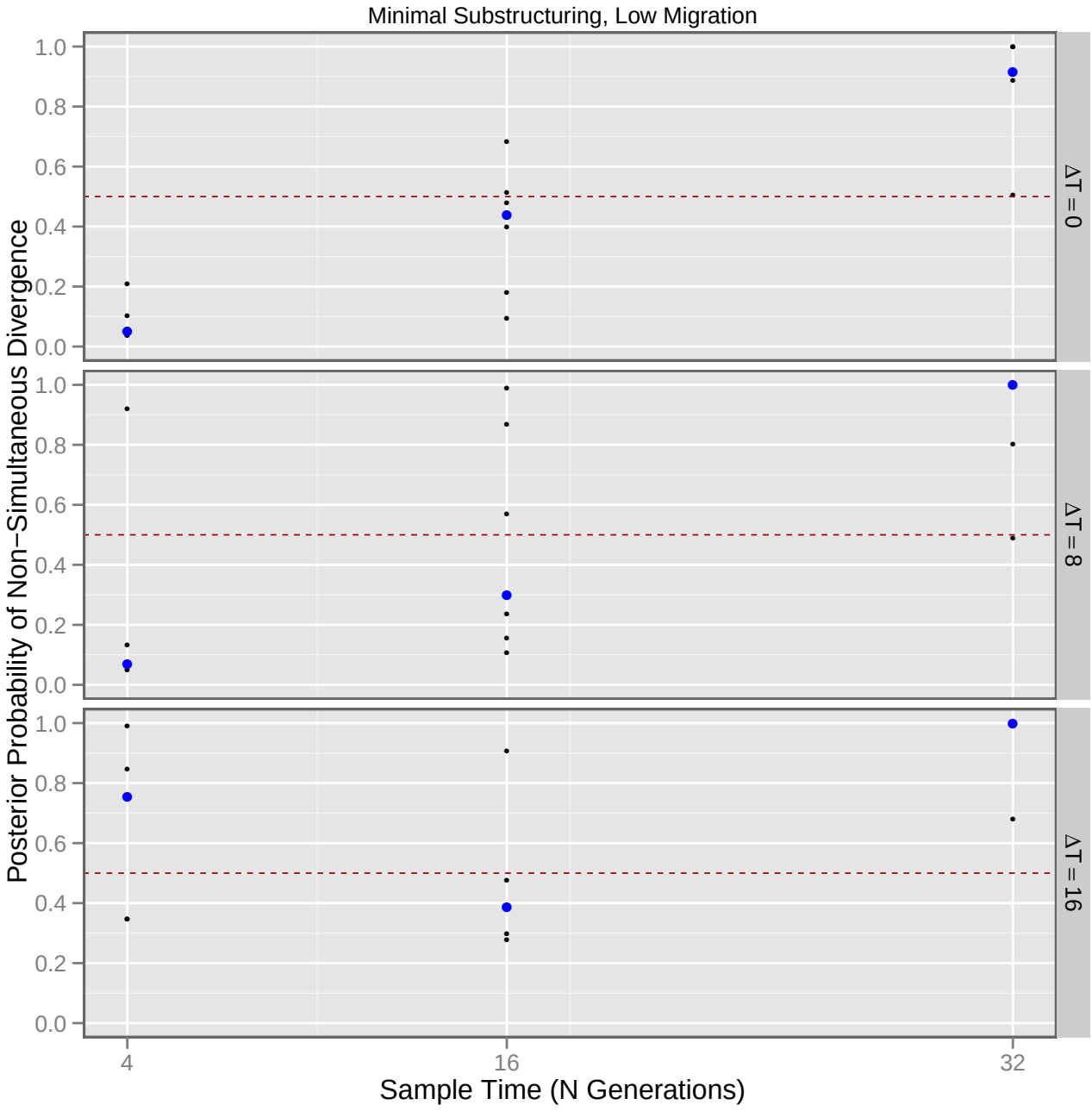


Figure 2.38: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values *for 5 independent loci* with *low* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

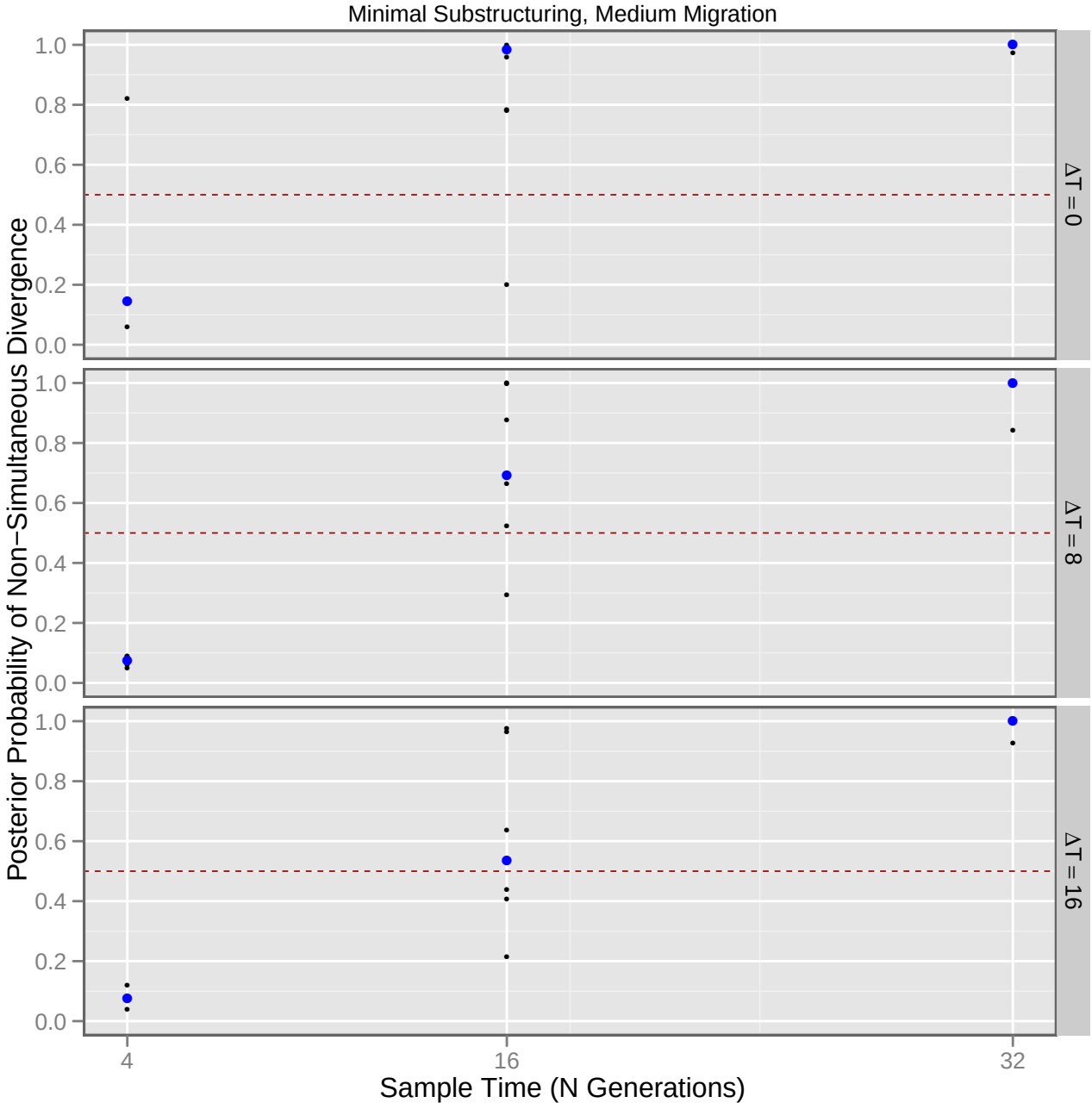


Figure 2.39: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values *for 5 independent loci* with *medium* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

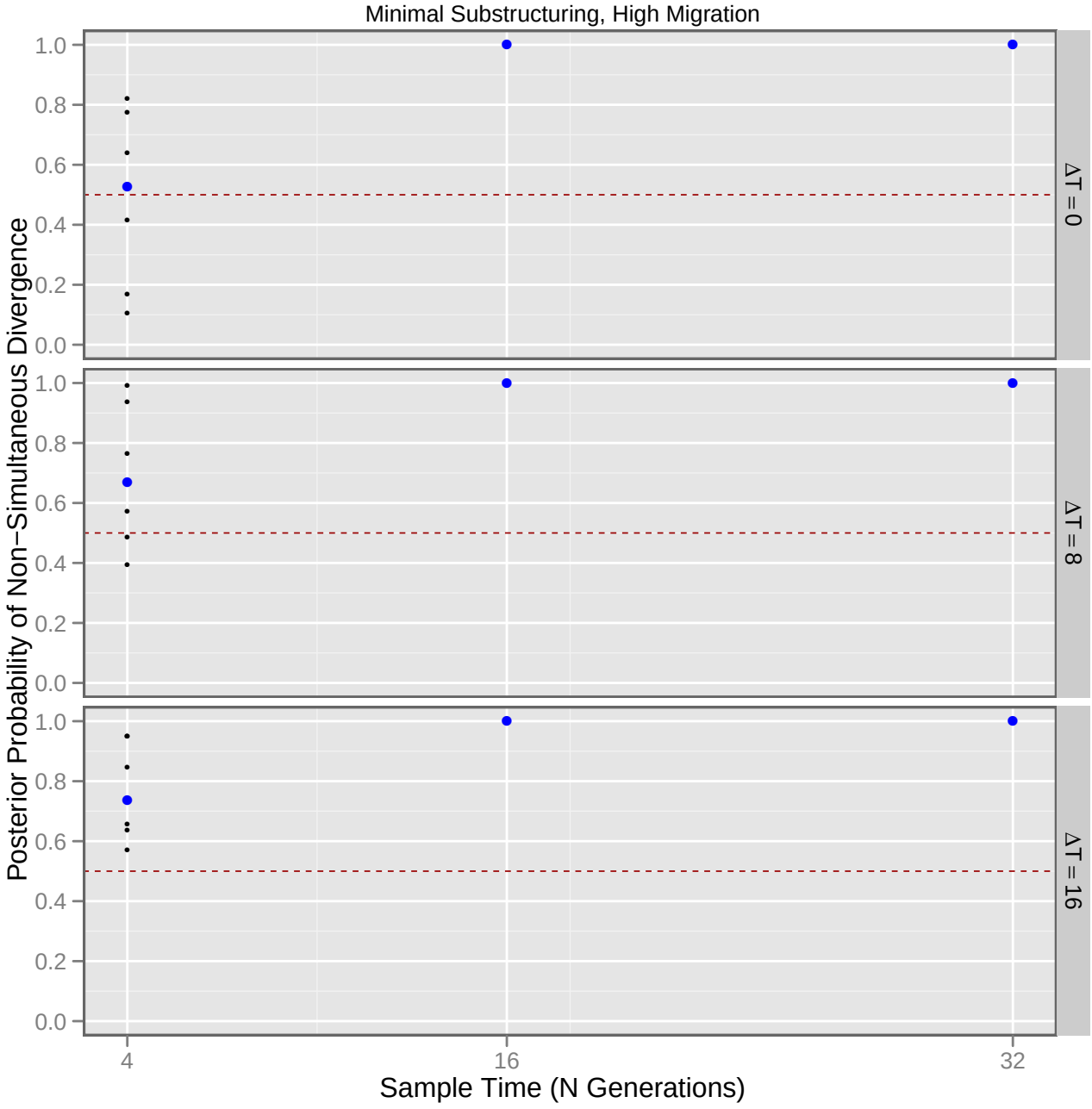


Figure 2.40: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values *for 5 independent loci* with *high* levels of post-vicariance gene flow when analyzed using `msbayes`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

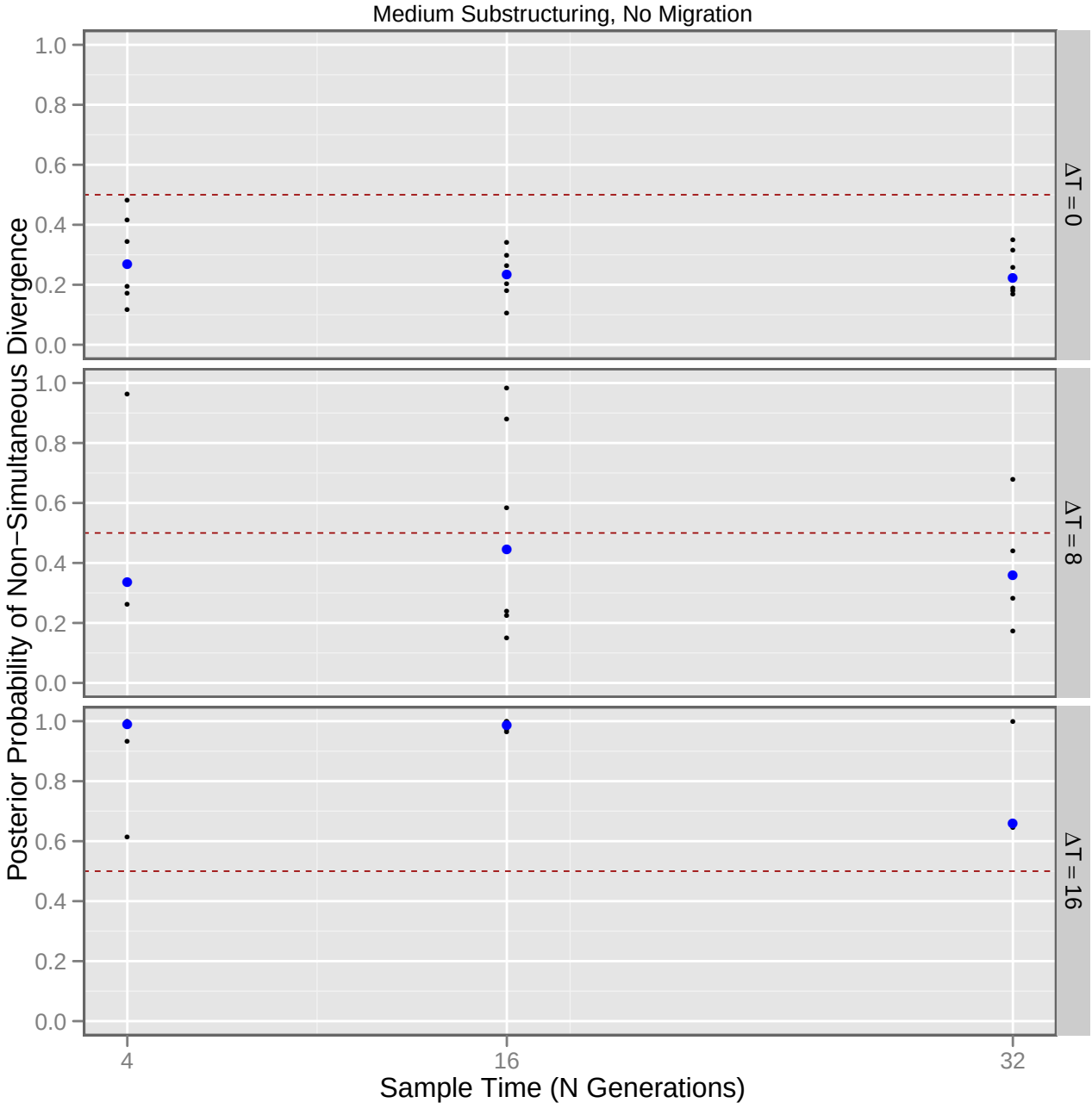


Figure 2.41: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values *for 5 independent loci* with *medium* levels of within population structuring when analyzed using `msbayes`. See figure 2.4 for details on interpreting the plots.

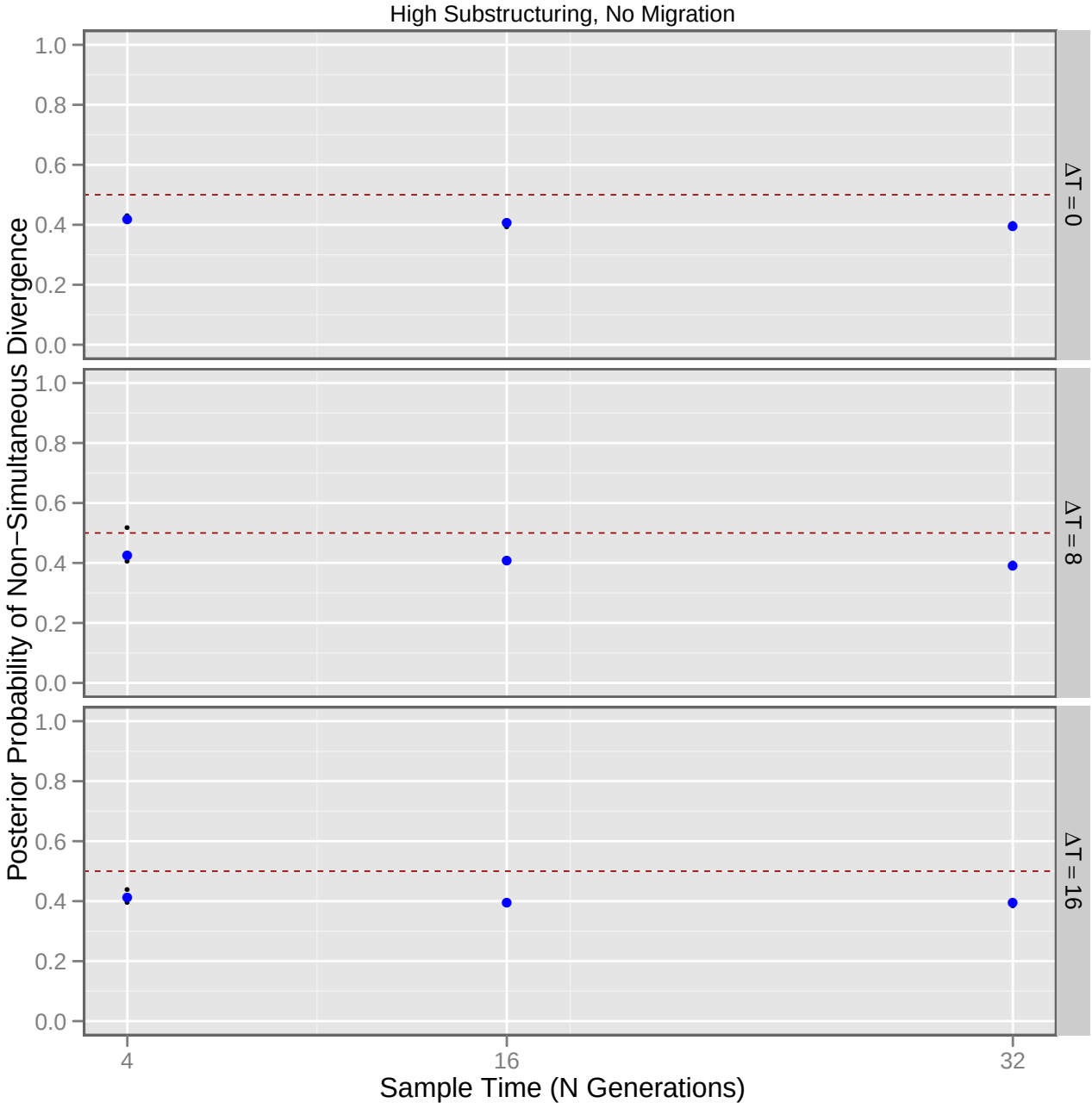


Figure 2.42: Posterior probability of non-simultaneous divergence of forward-time simulations under high theta values *for 5 independent loci* with *high* levels of within population structuring when analyzed using `msbayes`. See figure 2.4 for details on interpreting the plots.

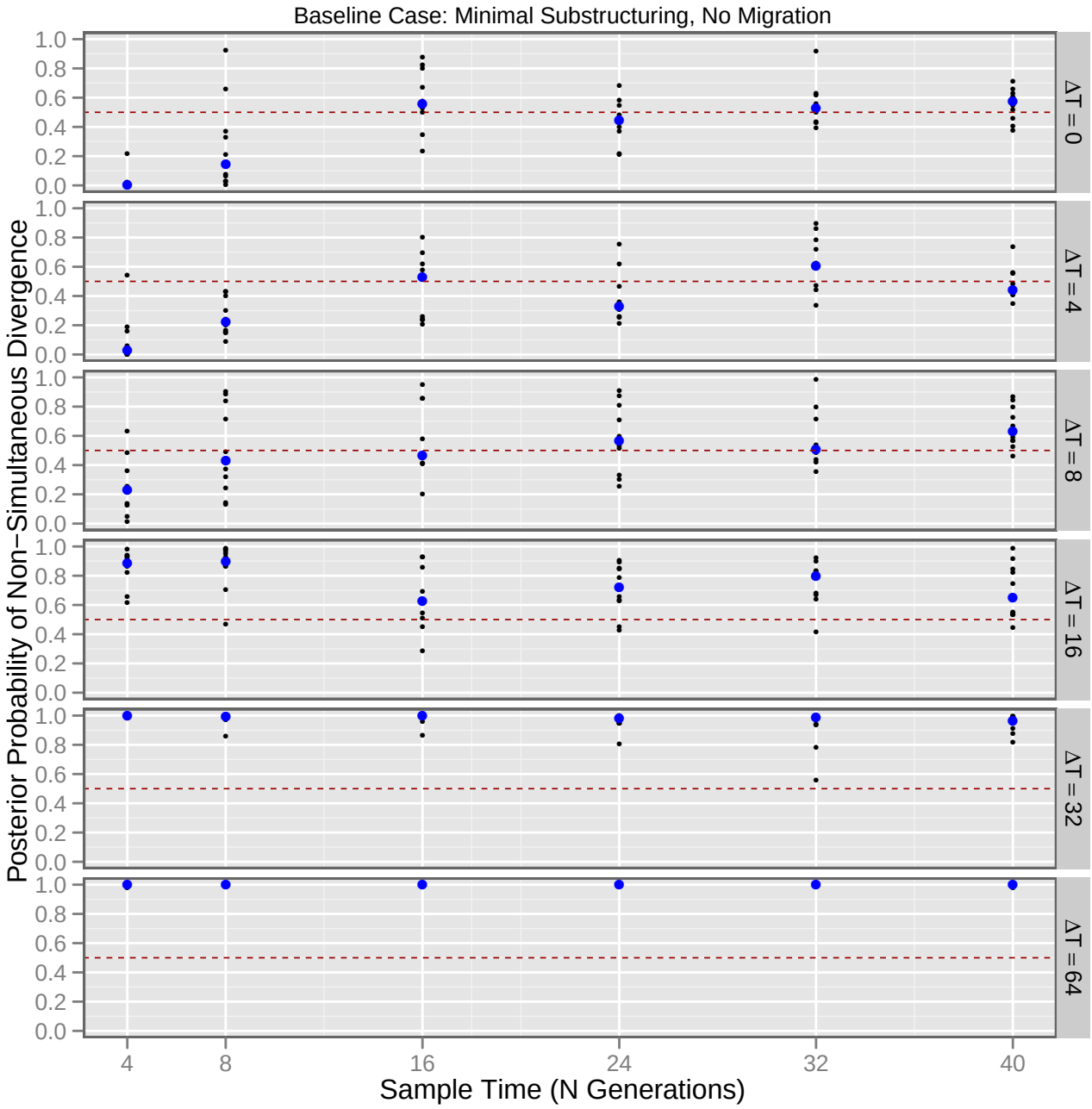


Figure 2.43: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under *low* theta values when analyzed using `IMa2`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.4 for details on interpreting the plots.

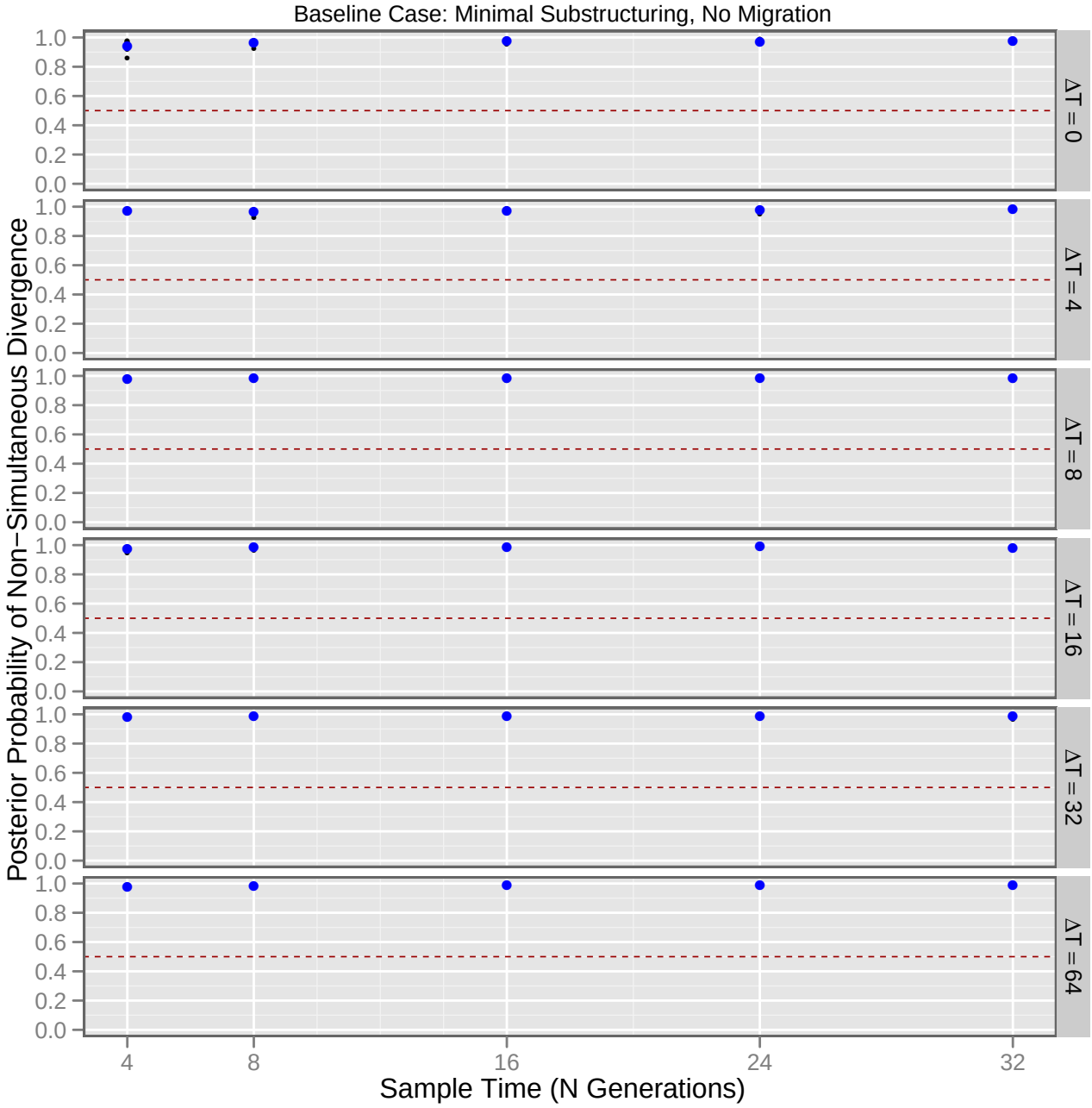


Figure 2.44: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under *high* theta values when analyzed using `IMa2`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 2.4 for details on interpreting the plots.

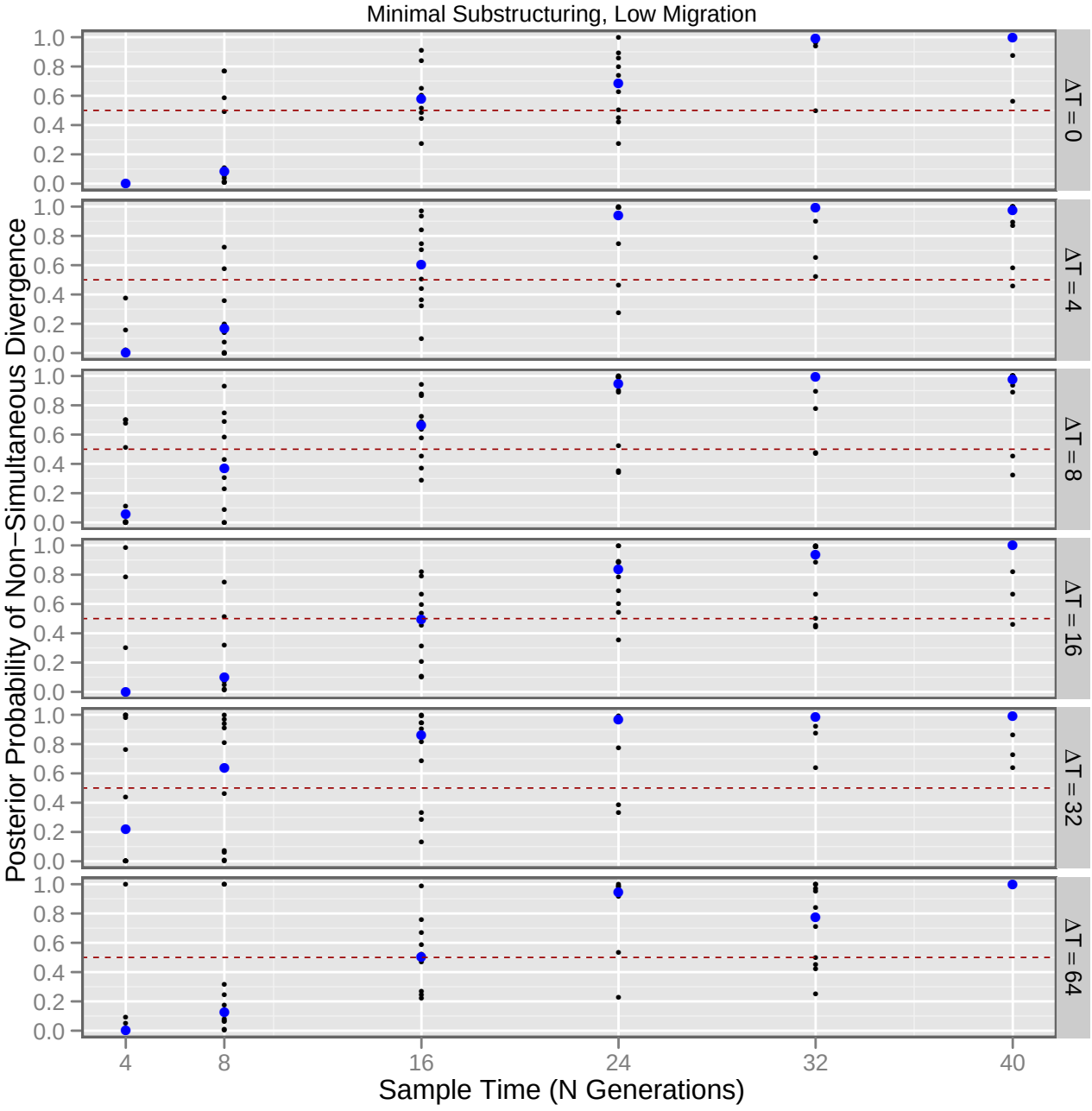


Figure 2.45: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *low* levels of post-vicariance gene flow when analyzed using `IMa2`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

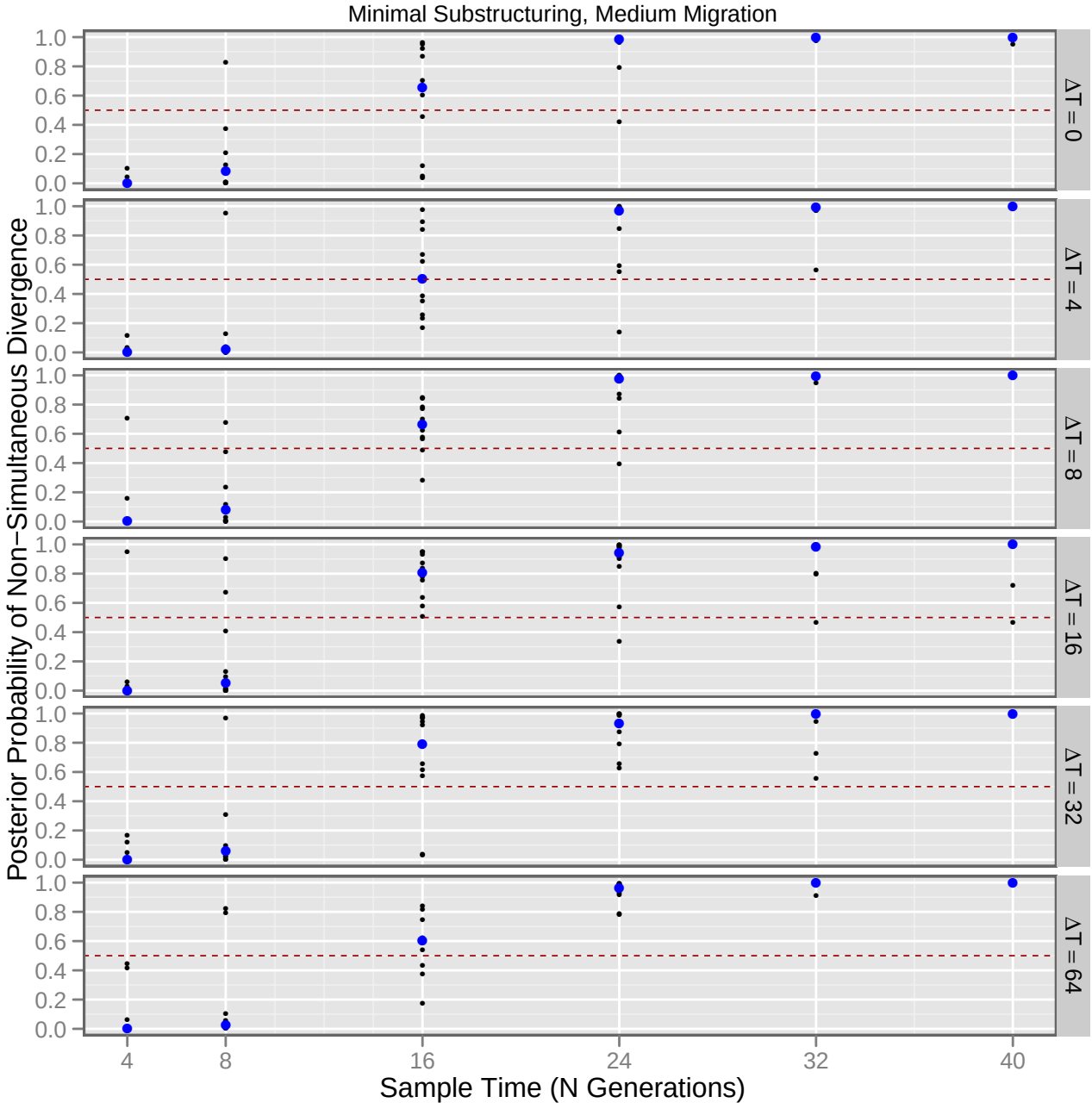


Figure 2.46: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *medium* levels of post-vicariance gene flow when analyzed using `IMa2`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

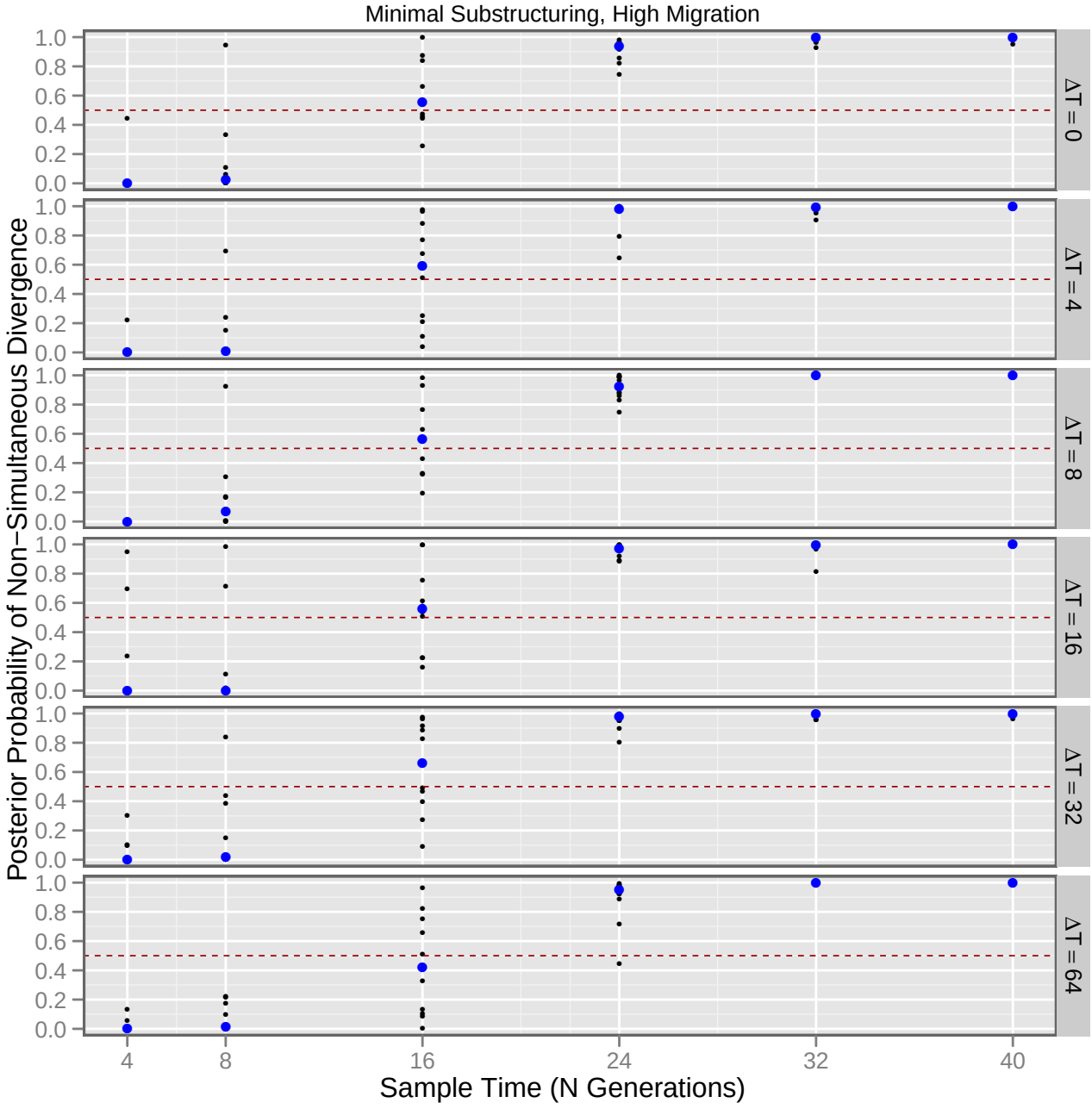


Figure 2.47: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *high* levels of post-vicariance gene flow when analyzed using `IMa2`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

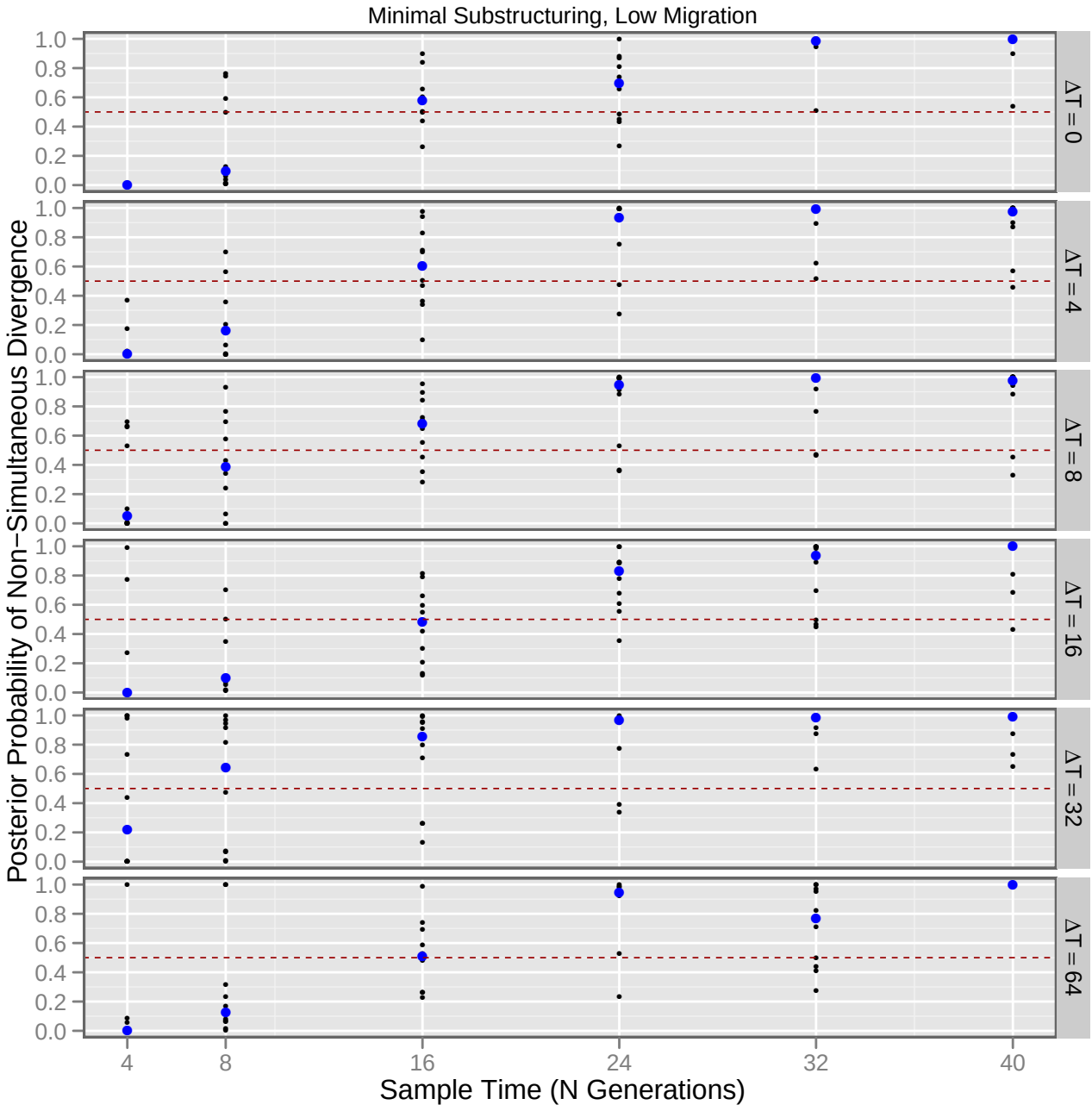


Figure 2.48: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *low* levels of post-vicariance gene flow when analyzed using `IMa2`, under an estimation model that *does* account for migration. See figure 2.4 for details on interpreting the plots.

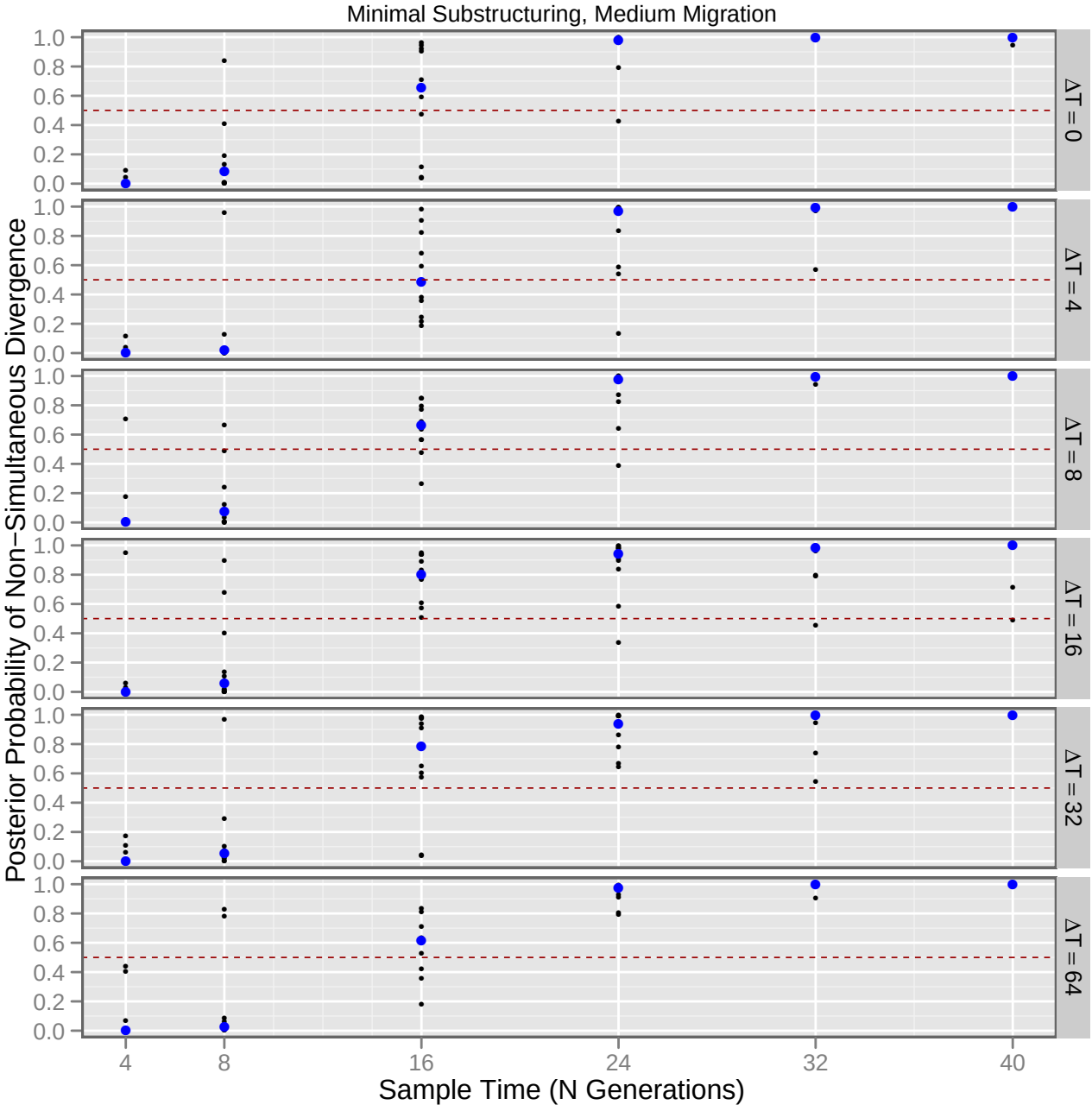


Figure 2.49: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *medium* levels of post-vicariance gene flow when analyzed using `IMa2`, under an estimation model that *does* account for migration. See figure 2.4 for details on interpreting the plots.

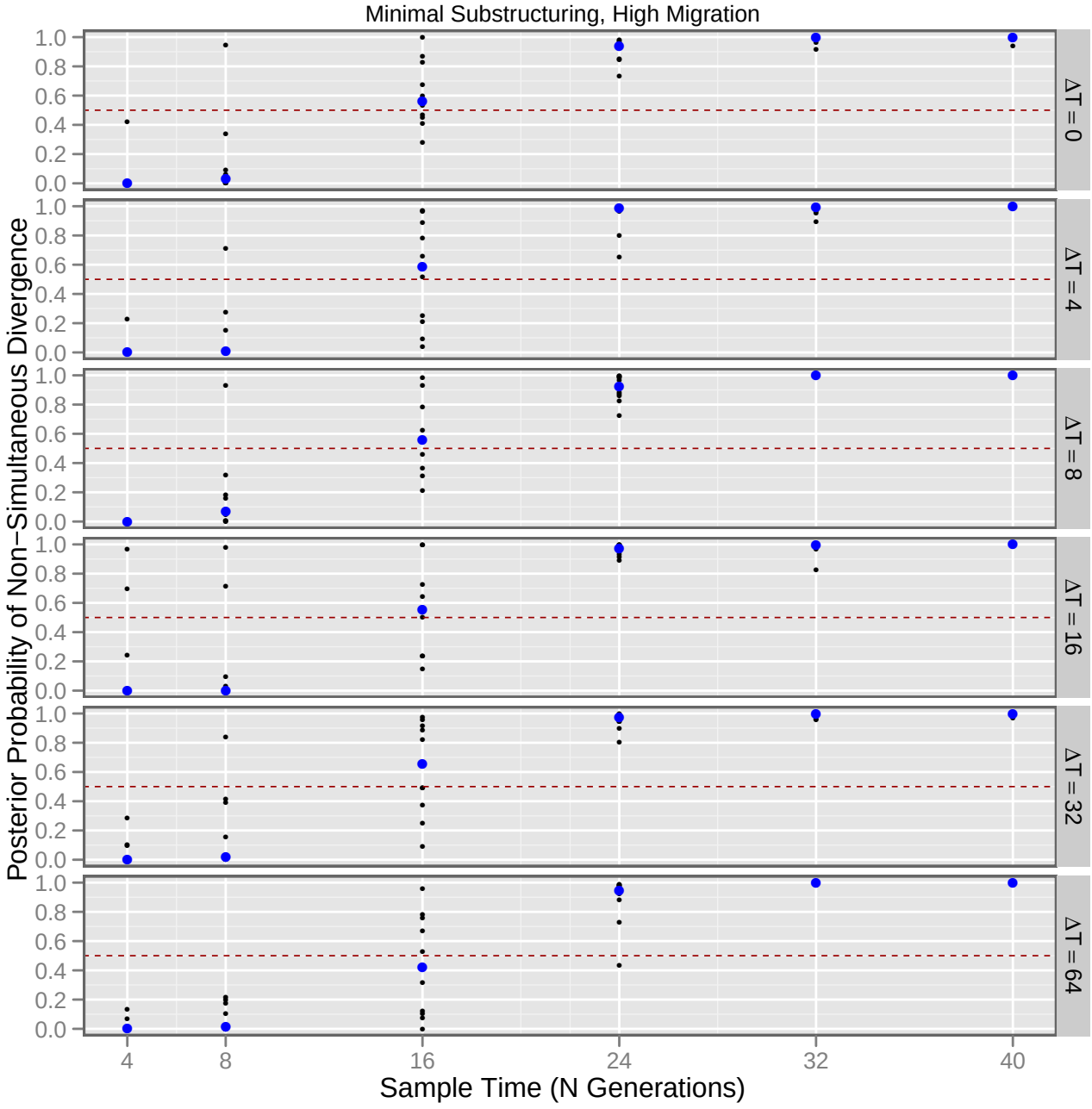


Figure 2.50: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *high* levels of post-vicariance gene flow when analyzed using `IMa2`, under an estimation model that does *not* account for migration. See figure 2.4 for details on interpreting the plots.

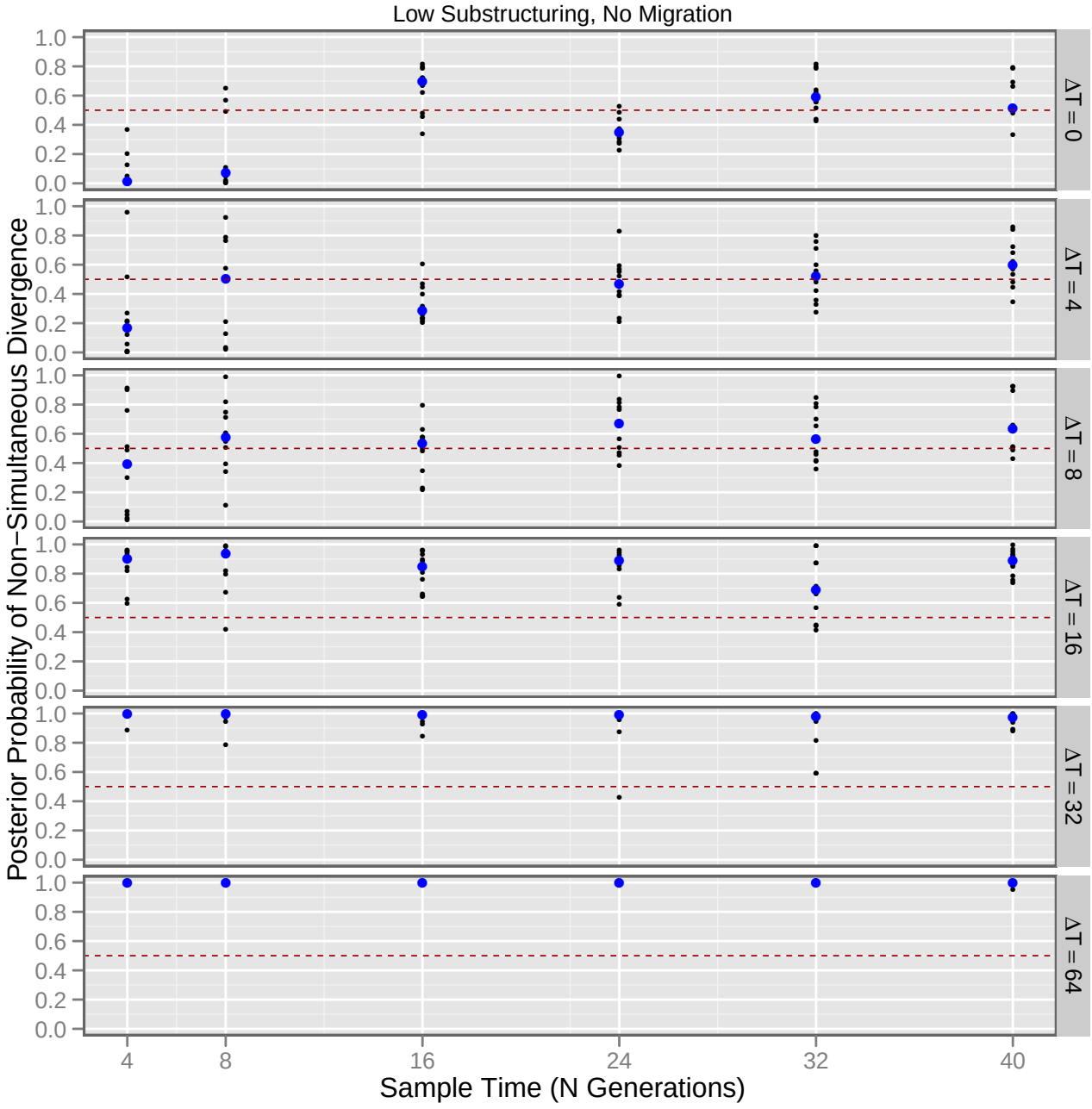


Figure 2.51: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *low* levels of within population structuring when analyzed using `IMa2`. See figure 2.4 for details on interpreting the plots.

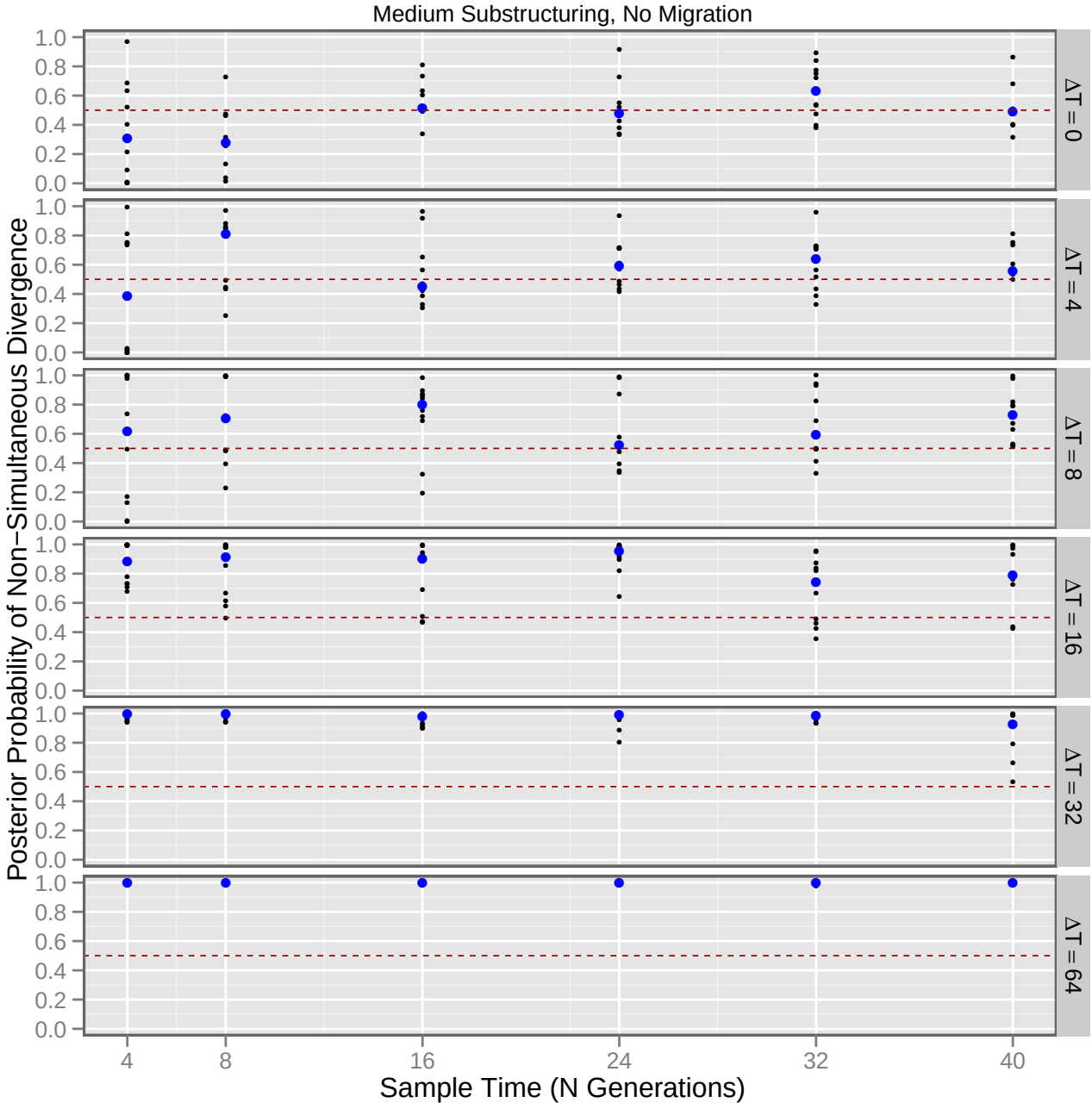


Figure 2.52: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *medium* levels of within population structuring when analyzed using `IMa2`. See figure 2.4 for details on interpreting the plots.

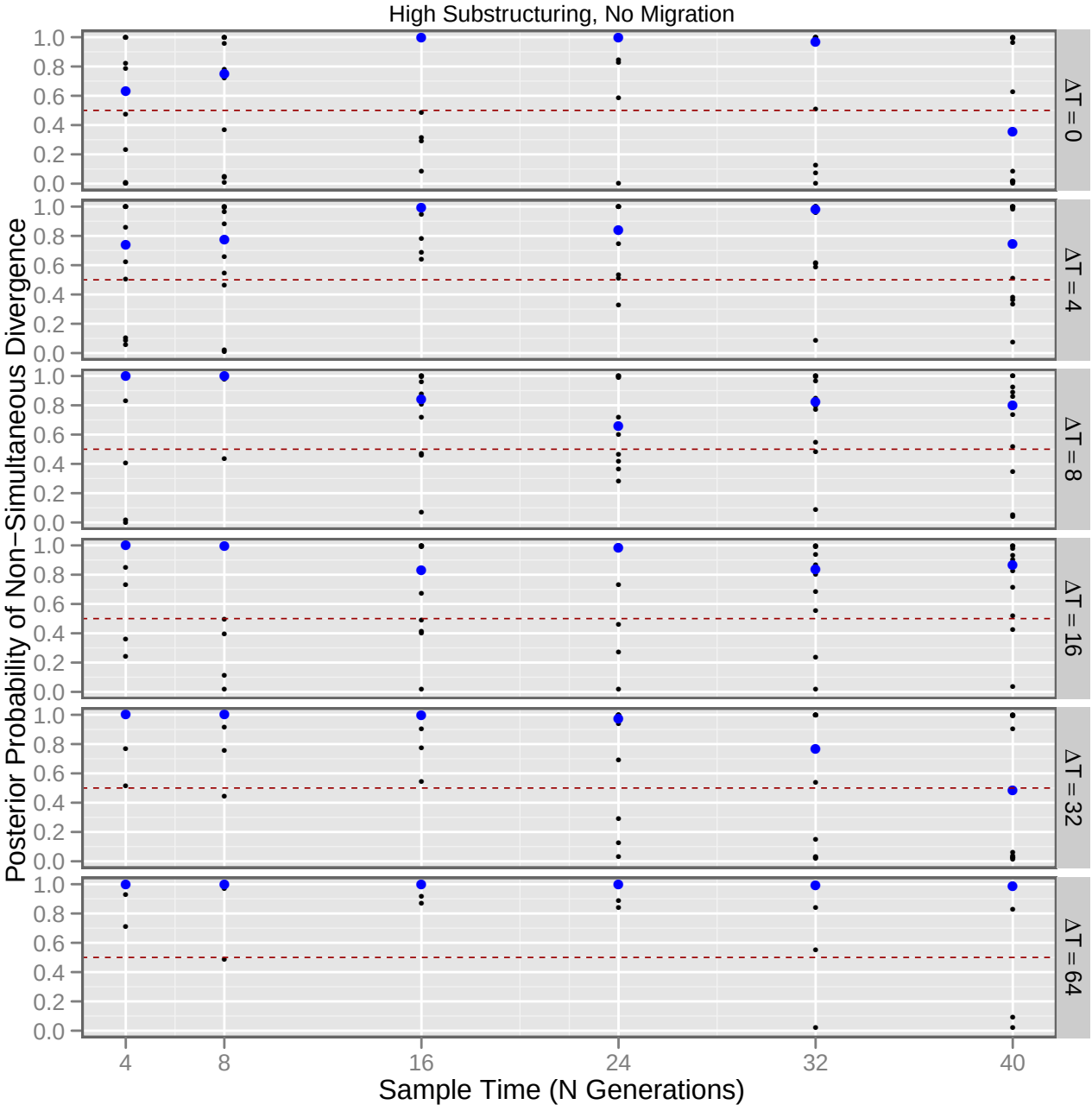


Figure 2.53: Posterior probability of non-simultaneous divergence of forward-time simulations under low theta values with *high* levels of within population structuring when analyzed using `IMa2`. See figure 2.4 for details on interpreting the plots.

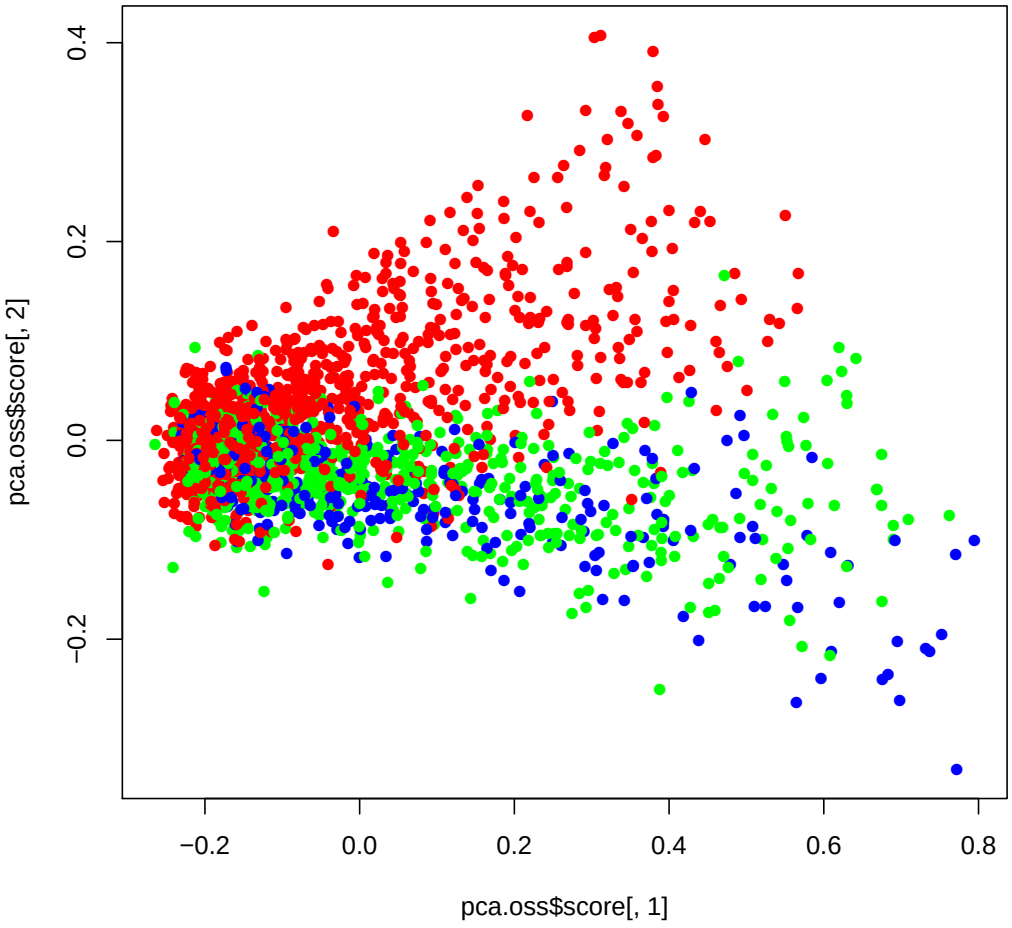


Figure 2.54: Plot of first two components resulting from a principal components analysis carried out on the summary statistics calculated across all single locus simulations. Blue dots represent true simultaneous divergence (i.e., $\psi = 1, \Delta T = 0$), green dots represent non-simultaneous divergence with $\Delta T = 4$ and $\Delta T = 8$, and red dots represent non-simultaneous divergence with $\Delta T \geq 16$.

Chapter 3

Full-Likelihood Bayesian Simultaneous Divergence Time Testing by Integrated Parallel Analysis of Multiple Genes of Multiple Species

3.1 Introduction

The previous chapter evaluated the performance of two different approaches to simultaneous divergence time testing. The first was an Approximate Bayesian Computation (ABC) approach, `msbayes` (Hickerson et al., 2006a; Huang et al., 2011), while the second was a full-likelihood Bayesian approach using `IMa2` (Hey, 2010).

It was found that even when all estimation model assumptions were met, the ABC approach only performed adequately within narrowly defined constraints on the conditions which generated the data: an upper limit on the time separating divergence events, a lower limit on the time elapsed since the divergence events, and a relatively higher range of mutation rates. If the data were sampled outside these constraints, the results produced were spurious and misleading, with support for incorrect conclusions and no indication that the method had actually failed. Furthermore, recognizing the data were sampled from outside these constraints required information that would

obviate or render uneccessary the analysis, In addition, the responses of the method when data were sampled outside these constraints and/or when the assumptions of the estimation model were violated were varied. In most cases, spurious support of varied degrees of strength were indicated for different models, depending on the nature of the deviation or violation of the constraints.

The most straightforward way more information can be added to an ABC analysis for simultaneous divergence time testing, and the only way feasible under real-world conditions, is the addition of data from more independent loci. However, the addition of up to five loci actually proved to degenerate performance, a conclusion corroborated by work by the `msbayes` authors (Huang et al., 2011), who found that when using up 16 loci, the performance of `msbayes` was worse relative to single locus data due to coalescence variation, and 32 loci or more were required for improvement in performance relative to single locus data.

The full-likelihood method using `IMa2` did not perform much better. Its performance envelope mirrored that of `msbayes`, except whereas `msbayes` required data generated under high-theta regimes and produced false support for the single divergence model when given data produced under low-theta regimes, the revese was case for `IMa2`, which performed adequately given data generated under relatively low-theta regimes but failed by producing support for *non*-simultaneous divergence when given data generated under relatively high-theta regimes.

This suite of performance and behavior characteristics of `msbayes` can be attributed to the summary statistics used: π, the mean pairwise differences between sequences; $\hat{\theta}_W$, Watterson's estimator of θ (Watterson, 1975); π_{net}, the difference between the mean pairwise differences of sequences *within* each daughter subpopulation and the mean pairwise differences of sequences *between* each daughter subpopulation; $\text{var}(\pi - \theta_W)$, the denominator of Tajima's D (Tajima, 1989), i.e. the variance of the difference between two difference estimates of the population mutation parameter θ; and Wakeley's ψ (Wakeley, 1996). It is possible that further study may lead to development of summary statistics that improve the behavior of `msbayes`. This view is challenged by the fact that the results produced under the full-likelihood method implemented using `IMa2` were not dramatically improved. It would be expected that if the reason for the limitations in the performance of `msbayes` were due to the loss of information in the summary statistics, a full-likelihood method would performance significantly better in contrast, as full likelihood-based coalescent approaches are the most powerful method we currently have. This might still be true. The problem with the

evaluation of the full-likelihood method using `IMa2` was that the estimation was run independently for each of the taxon pairs, and the numerical instability and loss of precision in the MCMC required large time bins — on the order of $16.67N$ or 41,675 generations. This was necessary because `IMa2` was really not designed for this analysis, and the results of interest in this evaluation were based on extracting parameters that `IMa2` was designed to integrate over rather than estimate. In principle, there is nothing in Bayesian theory to preclude usage of the `IMa2` statistical framework in this way. In practice however, the implementation of the `IMa2` program (e.g., the way it treated, recorded and reported the divergence times), resulted in the loss of much information and power. As a result, we do not really have an accurate assessment of the performance of a full-likelihood approach to simultaneous divergence time model selection to which to compare the `msbayes` approach.

The current study presents a full-likelihood Bayesian framework for the simultaneous (parallel) analysis of data from multiple loci of multiple species, co-estimating genealogical/phylogenetic, chronological as well as demographic parameters. As such, it allows for the direct assessment of the question of interest, under the most powerful statistical approaches we have available for phylogeographic model selection. Furthermore, because it uses a full likelihood approach, it is also not prone to the various issues that reduce the effectiveness or undermine Approximate Bayesian Computation approaches to parameter (such as the "curse of dimensionality" / bias toward the prior (Beaumont et al., 2002)), or model selection, such as Bayes factor inconsistencies depending on summary statistics used (Marin et al., 2011), unpredictable biases in model posterior probabilities (Robert et al., 2011), etc.

3.2 Statistical Framework

3.2.1 The Probability Model

Let $\mathbf{X}$ represent the data, i.e., the molecular sequences sampled from multiple loci of individuals from each of daughter population from each species, where $X_{i,j}$ is the alignment of the j^{th} locus sampled for the i^{th} species. If we let s be the number of species and a_i be the number of loci sampled for the i^{th} species, then $1 \leq i \leq s$, and $1 \leq j \leq a_i$. Each alignment $X_{i,j}$ consists of genes sampled from two daughter populations of species i, $P_{i,1}$ and $P_{i,2}$, which descended from an ancestral population $P_{i,0}$ that split T_i generations in the past. Let $\mathbf{T} = \{T_1, T_2, ...T_s\}$ be the vector

of times that the ancestral population of i^{th} species diverged into the two daughter populations $P_{i,1}$ and $P_{i,2}$ for all $i \in \{1, 2, ..., s\}$ (T_i also corresponds to the depth or age of the root split of the i^{th} species tree representation of the data). Let $G_{i,j}$ be an ultrametric phylogenetic tree, with edge lengths in units of generations, that explains alignment $X_{i,j}$ given mutation rate $\mu_{i,j}$. Let the collection of gene trees and mutation rates for the alignments in $\mathbf{X}$ be represented by the vector $\mathbf{G}$ and μ, respectively. Let $N_{i,j,k}$ be the population size for k^{th} population of the i^{th} species, $P_{i,k}$ ($0 \leq k \leq 2, 1 \leq i \leq s$), where $k = 1$ and $k = 2$ represent the daughter populations and $k = 0$ represents ancestral population. Let $\mathbf{N}$ be the vector of population sizes across all populations of all species. Then the posterior probability of the model, $\Pr(\mathbf{T}, \mathbf{G}, \mathbf{N})$ given the data is given by:

$$\Pr(\mathbf{T}, \mathbf{G}, \mathbf{N} \mid \mathbf{X}) = \frac{\Pr(\mathbf{X} \mid \mathbf{G}, \mu) \Pr(\mathbf{G} \mid \mathbf{N}, \mathbf{T}) \Pr(\mu) \Pr(\mathbf{N}) \Pr(\mathbf{T})}{\Pr(\mathbf{X})}. \tag{3.1}$$

The Likelihood

$\Pr(\mathbf{X} \mid \mathbf{G}, \mu)$ is the likelihood of the genealogies and mutation rate given the sequence data. For each species i, let $X_{i,j}$ represent the data (sequence alignment) for locus j, with genes sampled from both daughter populations, $P_{i,1}$ and $P_{i,2}$. Each genealogy $G_{i,j}$ then consists of a phylogenetic tree which relates the genes from across both populations of species i for locus j under a Jukes-Cantor finite states character substitution model with mutation rate $\mu_{i,j}$. This probability $\Pr(X_{i,j} \mid G_{i,j}, \mu_{i,j})$ is the phylogenetic or "Felsenstein" likelihood (Felsenstein, 1981, 2004; Yang, 2006), and the product of this is taken across all loci and species to yield $\Pr(\mathbf{X} \mid \mathbf{G}, \mu)$.

The Numerical Priors

$\Pr(\mu), \Pr(\mathbf{N})$, and $\Pr(\mathbf{T})$ are the joint prior distribution on the mutation rates, population sizes, and divergence times, respectively.

The Structured Coalescent Prior

$\Pr(\mathbf{G} \mid \mathbf{N}, \mathbf{T})$ is the joint structured coalescent prior on the genealogies, given by the structuring implied by the species tree divergence times, $\mathbf{T}$, and the population sizes associated with each edge of the species tree, $\mathbf{N}$. For any particular locus j of species i, the individual genes sampled from daughter populations $P_{i,1}$ and $P_{i,2}$ are related to each other by the genealogy $G_{i,j}$, as described

above. The genealogy $G_{i,j}$ is thus "embedded" within a two-tip species tree with root depth of T_i, which induces a mapping of the nodes in the genealogy onto one of the three edges of the species tree (the two sister descendent edges, each with an edge length of T_i, and the root edge). The leaf nodes of a particular genealogy are mapped to the species tree based on the populations from which the corresponding genes are sampled: $P_{i,1}$ or $P_{i,2}$. This identity is fixed and does not change, i.e. it is part of the data. All internal nodes of this genealogy are coalescence events, and are mapped onto a population based on the depth of genealogical node (i.e., the age of the node, or the time in generations from the present as given by the sum of edge lengths between the node and any one of the tips descended from it). Specifically, internal nodes of the genealogy with a depth $\leq T_i$ will be assigned to $P_{i,k}$ if and only if *both* of its children are from $P_{i,k}$. If the depth of an internal node is greater than T_i, then the node will be mapped to the ancestral population or root edge of the species tree, regardless of the population identities of its children (i.e., coalescence between genes of different populations are allowed in the ancestral species). If, on the other hand, the depth of an internal node is less than or equal to T_i on genealogy $G_{i,j}$, and its children are from *different* populations, then this implies a migration event, which is not supported under the current model, and in this special case, $\Pr(G_{i,j} \mid N_{i,k}, T_i) = 0$.

Thus given a genealogy $G_{i,j}$ relating genes sampled from two populations, $P_{i,1}$ and $P_{i,2}$ that diverged T_i generations ago, assuming that no migration is allowed, we can assign the internal nodes of the genealogy to one of three coalescence segments: one for each of the subpopulations $P_{i,1}$ and $P_{i,2}$, and one for the ancestral population of species i, $P_{i,0}$. Within each coalescence segments, the waiting times between coalescence events is given by the the differences in depth between each successive internal node if they are ranked in order of depths. Let $\omega_{i,j,k}$ be the vector of waiting times between coalescence events on genealogy $G_{i,j}$ for subpopulation $P_{i,k}$. The distribution of waiting times between coalescence events follows an exponential distribution with rate parameter of $\frac{\binom{n_t}{2}}{N_{i,k}}$ (Kingman, 1982a,b), where $n_{i,j,t}$ is the number of independent lineages remaining at time t in genealogy $G_{i,j}$ and $N_{i,k}$ is the (haploid) population size of $P_{i,k}$:

$$\Pr(\omega_{i,j,k} \mid N_{i,k}) = \prod_{h \in \omega_{i,j,k}} \frac{\binom{n_{i,j,t}}{2}}{N_{i,k}} e^{-\frac{\binom{n_{i,j,t}}{2}}{N_{i,k}} h}. \tag{3.2}$$

Further let $c_{i,j,k}$ be the number of *uncoalesced* lineages remaining in population $P_{i,k}$ at time $t = T_i$, for $k \in 1, 2$ (all lineages coalesce in the ancestral population $P_{i,0}$). The probability that none of the $c_{i,j,k}$ lineages coalesces in the each of the daughter edges between the time of the last coalescence (i.e., $d_{i,j,k}$, the depth of the deepest internal node on genealogy $G_{i,j}$ assigned to daughter population $P_{i,k}$) and the end of the edge is given by the law of total probability as the complement of the probability that none of the $c_{i,j,k}$ lineages colaesce in the remaining time interval, which is 1 minus the CDF of the exponential with a rate of $\frac{\binom{c_{i,j,k}}{2}}{N_{i,k}}$:

$$\begin{aligned}\Pr(c_{i,j,k} \mid d_{i,j,k}, N_{i,k}) &= 1 - (1 - e^{-\frac{\binom{c_{i,j,k}}{2}}{N_{i,k}} d_{i,j,k}}) \\ &= e^{-\frac{\binom{c_{i,j,k}}{2}}{N_{i,k}} d_{i,j,k}}\end{aligned} \tag{3.3}$$

Thus, for any particular genealogy for locus j of species i, $G_{i,j}$, given population sizes of $N_{i,k}$ for the two daughter ($k = 1$ and $k = 2$) and the ancestral ($k = 0$) populations and an ancestral population splitting time of T_i, the prior probability of the genealogy as given by the structured coalescent is:

$$\Pr(G_{i,j} \mid N_{i,k}, T_i) = \prod_{k=1,2} [\Pr(\omega_{i,j,k} \mid N_{i,k}) \Pr(c_{i,j,k} \mid d_{i,j,k}, N_{i,k})] \Pr(\omega_{i,j,0} \mid N_{i,0}), \tag{3.4}$$

assuming there there is no migration implied by any internal node in any of the daughter populations $P_{i,1}$ or $P_{i,2}$ having children from two differne populations. Otherwise, the $\Pr(G_{i,j}) = 0$. The prior probability for all genealogies, $\mathbf{G}$, is given by the product of the priors for each genealogy $G_{i,j}$ across all species and loci.

The Probability of the Data

The denominator in expression 3.1, $\Pr(\mathbf{X})$, is the probability of the data, and is given by the integration of the numerator across all parameters:

$$\Pr(\mathbf{X}) = \int_{\mathbf{T}} \int_{\mathbf{N}} \int_{\mu} \int_{\mathbf{G}} \Pr(\mathbf{X} \mid \mathbf{G}, \mu) \Pr(\mathbf{G} \mid \mathbf{N}, \mathbf{T}) \Pr(\mu) \Pr(\mathbf{N}) \Pr(\mathbf{T}) \, d\mathbf{G} \, d\mu \, d\mathbf{N} \, d\mathbf{T}. \tag{3.5}$$

Expression 3.5 is difficult to evaluate analytically, and this means that expression 3.1 cannot be evaluated directly. Here, instead of calculating the posterior using expression 3.1 directly, we use Metropolis-Hastings Markov chain Monte Carlo (MCMC) to sample from the posterior distribution (Metropolis et al., 1953; Hastings, 1970).

3.2.2 Evaluating the Posterior Using Metropolis-Hastings Markov chain Monte Carlo

MCMC is a class of algorithms designed to sample from a target density, which, in this case is the posterior given in expression 3.1. Assuming that we have initialized the MCMC chain to a starting state, for the next and every subsequent step, we propose a state which we accept as the new current state with a probability equal to the ratio of the posterior of the proposed state to the current state (or 1, if this ratio is greater than 1).

That is, if we let θ represent the current state of the system, and θ^* represent the proposed state of the system (where "state" indicates the vector of values for $\mathbf{T}$, $\mathbf{G}$, $\mathbf{N}$, and μ), then the probability of accepting the proposed state using the algorithm of Metropolis et al. (1953) is:

$$\alpha = \min\left(1, \frac{\Pr(\mathbf{X} \mid \theta^*)}{\Pr(\mathbf{X} \mid \theta)} \frac{\Pr(\theta^*)}{\Pr(\theta)}\right). \tag{3.6}$$

The aspect of MCMC that allows it to be used to evaluate functions that cannot be solved analytically is that the value of interest is a ratio of the function evaluated at the proposed and current states. In the context of sampling from the posterior in a Bayesian analysis, this means that the denominator on the right-hand side of expression 3.1, i.e., the probability of the data, cancels out, and thus at every MCMC step, only the numerator on the right hand side of expression 3.1

needs to evaluated:

$$\Pr(\mathbf{X} \mid \mathbf{G}, \mu) \Pr(\mathbf{G} \mid \mathbf{N}, \mathbf{T}) \Pr(\mu) \Pr(\mathbf{N}) \Pr(\mathbf{T}) \tag{3.7}$$

The original algorithm as formulated by Metropolis et al. (1953) assumed symmetrical proposals, where the probability of proposing a new state, θ^* from the current state, θ, is equal to the probability of proposing state θ from the θ^*. That is, equation 3.6 assumes that $q(\theta^* \mid \theta) = q(\theta \mid \theta^*)$, where $q(x \mid y)$ represents the probability of proposing state x from state y. This assumption was relaxed when the algorithm was extended by Hastings (1970) to allow for assymetrical proposals by factoring in a ratio, the Metropolis-Hastings ratio, into the acceptance probability. That is if the proposal densities are not constrained to be equal, the acceptance probability α is given by:

$$\alpha = \min\left(1, \frac{\Pr(\mathbf{X} \mid \theta^*)}{\Pr(\mathbf{X} \mid \theta)} \frac{\Pr(\theta^*)}{\Pr(\theta)} \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)}\right). \tag{3.8}$$

The collection of states visited at every step forms the MCMC chain. If the chain is *irreducible* (i.e., under the proposal density $q(\cdot \mid \cdot)$, it should be possible for the chain to reach any state from any other state given enough time), and all the states in the chain are *ergodic* — that is, every state is positive recurrent (i.e., with infinite time, each and every state should be visited an infinite number of times) and aperiodic (i.e., there is no particular constant/fixed number of steps between every visit to any given state) — then the chain itself is said to be ergodic. If a chain is ergodic, then from any starting state, given enough time, the chain will converge on the target density such that the sampling frequency of states in chain approximate the sampling frequency of states from the target density, regardless of the starting state of the chain (i.e. an ergodic chain will converge to a stationary distribution that approximates the target distribution regardless of the starting state).

3.2.3 MCMC Moves

Construction of the MCMC chain proceeds using a collection of different moves or proposal types, each of which perturbs a particular component or subset of components of the state as expressed

in equation 3.7. The moves currently implemented are:

- Gene subtree sliding move.
- Divergence time move.
- Mutation rate move.
- Population size move.
- Divergence time model jump move.

Gene Subtree Sliding Move

In this move, a non-root edge is randomly selected from a random genealogy, and slided up or down (toward the root or tip) by a random amount, adjusting the length of the edge to preserve the ultrametricity of the tree. Let e be a non-root edge selected with uniform random probability from all non-root edges on genealogy, $G_{i,j}$ selected with uniform random probability from all gene trees. Let the current depth of the parent node of e be d_0. Let the depth of the child nodes of the parent node of e be d_1 and d_2 for the first and second child, respectively. Let w be a tuning parameter that determines the maximum magnitude of depth displacement in a gene subtree sliding move. Then the new depth for the parent node of e, d_0^* is given by:

$u_1 \leftarrow Uniform(0, 1)$
if $u_1 \geq 0.5$ **then**
  $d_0^* \leftarrow d_0 + Uniform(0, w)$
else
  $d_0^* \leftarrow d_0 - Uniform(0, \min(w, d_0 - \max(d_1, d_2)))$
end if

Basically, a direction for the slide, either upwards toward the root or downward toward the tip, is determined with uniform random probability. If sliding upwards, then the depth displacement is simply elected with uniform random probability from a range given by $[0, w)$. Sliding downwards toward the tip, however, is a little more complicated, as the state space does not allow for the depth of a node to be less (i.e., younger) than the depth of any of its children. Hence the upper bound of the depth displacement must be constrained by both the window size *and* the difference in depths between the parent node of the edge being slided and depths of its children (it might be

possible for the difference in depth between the parent node of the edge being slided depths of the children to be greater than the window size; the window size constraint is to ensure that the move is reversible to satisfy the irreducible constraint on the chain).

As a result of the special constraints of sliding downward, the probability associated from sliding down from a particular depth x to any other particular depth x' is not equal to the probability of sliding up from x' to x. Furthermore, when a subtree is slided up toward the root and crosses parent nodes, there is only one path to choose. However in the revese case, when a subtree is being slided toward the tip, it needs to randomly decide which subpath (left or right) it needs to slide down. Again, this results in an unequal probability between a move and its reverse move. Both of these unequalities must be taken into account in the Hastings ratio for the proposal. If the number of nodes traversed when sliding is n_c, then the Hastings ratio for a downward slide is given by:

$$2^{n_c} \frac{\min(w, d_0 - \max(d_1, d_2))}{w}, \tag{3.9}$$

while the Hastings ratio for an upward slide is given by:

$$0.5^{n_c} \frac{w}{\min(w, d_0 - \max(d_1, d_2))}. \tag{3.10}$$

The acceptance probability for a tipward slide move is thus given by in minumum of 1 and:

$$\frac{\Pr(X_{i,j} \mid G^*_{i,j}, \mu_{i,j})}{\Pr(X_{i,j} \mid G_{i,j}, \mu_{i,j})} \frac{\Pr(G^*_{i,j} \mid T_i, \mathbf{N_i})}{\Pr(G_{i,j} \mid T_i, \mathbf{N_i})} \left(\frac{2^{n_c} \min(w, d_0 - \max(d_1, d_2))}{w} \right). \tag{3.11}$$

while that for a rootward slide move is given by the minimum of 1 and:

$$\frac{\Pr(X_{i,j} \mid G^*_{i,j}, \mu_{i,j})}{\Pr(X_{i,j} \mid G_{i,j}, \mu_{i,j})} \frac{\Pr(G^*_{i,j} \mid T_i, \mathbf{N_i})}{\Pr(G_{i,j} \mid T_i, \mathbf{N_i})} \left(\frac{0.5^{n_c} w}{\min(w, d_0 - \max(d_1, d_2))} \right). \tag{3.12}$$

Divergence Time Move

In this move, a species i is selected at random the depth of the root split T_i is displaced a value drawn in uniform probability over the range $[-w, w]$, where w is a tuning parameter specifying magnitude of the maximum possible depth offset:

$u_1 \leftarrow Uniform(0, 1)$
if $u_1 \geq 0.5$ **then**
  $T_i^* \leftarrow T_i + Uniform(0, w)$
else
  $T_i^* \leftarrow T_i - Uniform(0, w)$
end if

The Hastings ratio for this move is 1.0. The acceptance probability this move is thus simply given by the minimum of 1 and the ratio of the posteriors:

$$\frac{\Pr(\mathbf{G_i} \mid T_i^*, \mathbf{N_i})}{\Pr(\mathbf{G_i} \mid T_i, \mathbf{N_i})} \frac{\Pr(T_i^*)}{\Pr(T_i)}. \tag{3.13}$$

Mutation Rate Move

In this move, the mutation rate for a gene tree selected at random, $G_{i,j}$ is displaced with a value drawn with uniform probability over the range $[-w, w]$, where w is a tuning parameter specifying magnitude of the maximum possible mutation rate change.

$u_1 \leftarrow Uniform(0, 1)$
if $u_1 \geq 0.5$ **then**
  $\mu_{i,j}^* \leftarrow \mu_{i,j} + Uniform(0, w)$
else
  $\mu_{i,j}^* \leftarrow \mu_{i,j} - Uniform(0, w)$
end if

The Hastings ratio for this move is 1.0. The acceptance probability this move is thus simply given by the minimum of 1 and the ratio of the posteriors:

$$\frac{\Pr(X_{i,j} \mid G_{i,j}\mu_{i,j}^*)}{\Pr(X_{i,j} \mid G_{i,j}\mu_{i,j})} \frac{\Pr(\mu_{i,j}^*)}{\Pr(\mu_{i,j})}. \tag{3.14}$$

Population Size Move

In this move, a population $P_{i,k}$ is selected at random from a species i selected at random, and the population size is displaced by a value drawn with uniform probability over the range $[-w, w]$, where w is a tuning parameter specifying magnitude of the maximum possible size change.

$u_1 \leftarrow Uniform(0, 1)$
if $u_1 \geq 0.5$ **then**
  $N^*_{i,k} \leftarrow N_{i,k} + Uniform(0, w)$
else
  $N^*_{i,k} \leftarrow N_{i,k} - Uniform(0, w)$
end if

The Hastings ratio for this move is 1.0. The acceptance probability this move is thus simply given by the minimum of 1 and the ratio of the posteriors:

$$\frac{\Pr(\mathbf{G_i} \mid T_i, \mathbf{N^*_i})}{\Pr(\mathbf{G_i} \mid T_i, \mathbf{N_i})} \frac{\Pr(\mathbf{N^*_i})}{\Pr(\mathbf{N_i})} \tag{3.15}$$

Divergence Time Model Jumping Move

A model jump move is one in which the dimensionality of the chain is changed. Green (1995, 2003) describes a general procedure for changing model dimensionality such that the MCMC chain generates samples from the joint distribution of parameters and model indices. In the current implementation, we allow for a two-species system in which the divergence times, T_1 and T_2, are allowed to vary independently (the "split state"), as well as when they are constrained to be equal (the "merged state").

A move from the split state to the merged state involves making the older of the divergence times of the two species equal to the younger:

if $T_1 > T_2$ **then**
  $T_1 \leftarrow T_2$
else
  $T_2 \leftarrow T_1$
end if

A move from the merged state to the split state, on the other hand involves selecting one of

the two species, i, and making the depth of its root split, T_i, older by drawing a new depth with a uniform probability over a range bounded on the lower end by the current depth and the upper end by z, the depth of the youngest internal node of the collection of gene trees for species i that is the parent of child nodes from different populations.

$u_1 \leftarrow Uniform(0,1)$
if $u_1 < 0.5$ **then**
  $i \leftarrow 1$
else
  $i \leftarrow 2$
end if
$T_i^* \leftarrow T_i + Uniform(0, z - T_i)$

In the procedure described by Green (2003), the Hastings ratio is replaced by:

$$\frac{g'(\mathbf{u}')}{g(\mathbf{u})}|J|, \tag{3.16}$$

where:

- $\mathbf{u}$ is the set of random numbers generated using a probability distribution with the joint probability density $g(\mathbf{u})$, and deterministically results in the proposed state when used in the proposal function, $\theta^* = h(\theta, \mathbf{u})$;
- $\mathbf{u}'$ is a set of random numbers generated with a joint probability density $g'(\mathbf{u}')$, and deterministically produces the reverse move, i.e., the current state given the proposed state, $\theta = h(\theta^*, \mathbf{u}')$;
- the sum of sizes of dimensionality of the proposed states and the set of random numbers required for the reverse should equal the sum of the sizes of the dimensionality of the current state and the set of random numbers required for the forward move: $|\theta^*| + |\mathbf{u}'| = |\theta| + |\mathbf{u}|$,
- and $|J|$ is the absolute value of the determinant of the Jacobian of the transformation from $\{\theta, \mathbf{u}\}$ to $\{\theta^*, \mathbf{u}'\}$, $J = det\left[\frac{\delta(\theta^*, \mathbf{u}')}{\delta(\theta, \mathbf{u})}\right]$.

In our current implementation, the merge move involves going from a model with two independent divergence times to a single divergence time. Hence the pre-move state space has one extra

dimension. Without loss of generality, we let t_1 equal the younger of the two divergence times and t_2 equal the older: $t_1 = \min(T_1, T_2)$ and $t_2 = \max(T_1, T_2)$. Then, if θ represents the two-divergence time model state and θ^* the single-divergence time model state, $\theta = \{t_1, t_2\}$ and $\theta^* = \{t_1\}$. Furthermore, the merge move is deterministic, so $\mathbf{u} = \emptyset$, while the reverse move requires one random variable, the new divergence time of the species that is split away from the merged group, t_2, so that $\mathbf{u}' = \{t_2\}$. The Jacobian is then:

$$\begin{aligned} J &= det \begin{bmatrix} \frac{\partial \theta_1^*}{d\theta_1} & \frac{\partial \theta_1^*}{d\theta_2} \\ \frac{\partial u_1'}{d\theta_1} & \frac{\partial u_1'}{d\theta_2} \end{bmatrix} \\ &= det \begin{bmatrix} \frac{\partial (t_1)}{dt_1} & \frac{\partial (t_1)}{dt_2} \\ \frac{\partial (u_1')}{dt_1} & \frac{\partial (u_1')}{dt_2} \end{bmatrix} \\ &= det \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\ &= 1. \end{aligned} \tag{3.17}$$

The absolute value of the Jacobian is thus $|J| = |1| = 1$.

The ratio of the joint probability density of the random variables in the reverse move to that of the forward move is

$$\begin{aligned} \frac{g(\mathbf{u}')}{g(\mathbf{u})} &= \frac{\frac{1}{z}}{1} \\ &= \frac{1}{z}. \end{aligned} \tag{3.18}$$

In addition, we also need to account for the difference in proposal probabilities: while the merge is deterministic, there are two possible ways to split a merged group (i.e., either one of the pair of species may be selected to have its divergence time made older), and thus the ratio of proposing the reverse (split) move to that of the forward (merge) move is:

$$\frac{\frac{1}{2}}{1} = \frac{1}{2}. \tag{3.19}$$

Thus the complete Hastings ratio for the merge move is given by:

$$\begin{aligned} \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)} &= \frac{1}{2}\frac{g'(\mathbf{u}')}{g(\mathbf{u})}|J| \\ &= \frac{1}{2z}. \end{aligned} \tag{3.20}$$

The Hastings ratio for the reverse move is given by the reciprocal of expression 3.20.

The acceptance probability of the merge move is then the minimum of 1 and:

$$\frac{\Pr(\mathbf{G_1} \mid T_1)}{\Pr(\mathbf{G_1} \mid T_1)} \frac{\Pr(\mathbf{G_2} \mid T_1)}{\Pr(\mathbf{G_2} \mid T_2)} \frac{\Pr(T_1)}{\Pr(T_1)\Pr(T_2)} \frac{1}{2z}, \tag{3.21}$$

while the acceptance probability of the split move is the reciprocal of this.

3.3 Implementation

Estimation under the model and the MCMC scheme described above was implemented in a C++ program, "`beluga`". Calculation of the phylogenetic or "Felsenstein" likelihood was delegated to the BEAGLE library (Ayres et al., 2011), using the "pytbeaglehon" wrapper (Holder, 2011). Input is in the form of a set of NEXUS-formatted files (Maddison et al., 1997), with one file per species, with the file reading supported by the Nexus Class Library (Lewis, 2003). The current implementation requires the starting gene trees to be specified, in addition to the alignment and population identities of the sampled taxa.

A custom NEXUS block allows for specification of priors, parameter linkages as well as MCMC tuning parameter values. In addition to common probability distributions, such as the uniform or exponential, a fixed value can be assigned to priors to reflect existing knowledge or for debugging purposes. Moves can be assigned different relative weights, allowing for some moves to be proposed

more often than others, or to be disabled completely.

The primary output is a tab-delimited text file of parameters as well as a series of tree files, one tree file per locus, consisting of samples from the MCMC chain taken at user-specified intervals. Supporting scripts make use of the DendroPy phylogenetic computing library (Sukumaran and Holder, 2010) to help compose the input data files as well as summarize or analyze results.

3.4 Validation

3.4.1 Methods

Validation of Phylogenetic Likelihood Calculations

Unit testing of the phylogenetic likelihood calculations was based on comparing likelihoods calculated on test datasets to the results obtained in PAUP* (Swofford, 1998). A variety of datasets, both single locus and multi-locus, were used, and calculations were tested both of "clean" input trees as well as trees that had undergone several thousand rounds of MCMC transformations.

Validation of Gene Subtree Sliding Move

Given a null alignment, a divergence time or root split of 0, and a fixed population size of N_y, we expect that a well-behaved MCMC chain should sample gene trees with characteristics of neutral coalescent trees sampled from a population of size N_y. This is because the null alignment provides no information to the gene trees, and they are then essentially being sampled from the prior. Because the species root split is fixed at a depth of 0, the prior on the gene tree is that of an unstructured coalescence taking place in the ancestral population, which has a haploid population size of N_y. Thus the gene trees sampled from the posterior should be expected to be sampled from a coalescence distribution with a population size of N_y.

We carried out a series of tests to verify that `beluga` conformed to these expectations, using population sizes of 17,000 and 49,000, and gene trees with 20 leaves. A total of 10 replicates was carried out in each test, with the MCMC chain run for 1,000,000 steps and sampled every 10,000 steps. The coalescence interval sizes (i.e., difference between successive depths of internal nodes) along with the corresponding number of lineages existing at the time of coalescence were extracted

from each gene tree sampled and stored. The coalescence interval sizes for each particular number of existing lineages were pooled. The distribution of coalescence interval sizes given a particular number of existing lineages, $k, k \in \{2, 3, ..20\}$, should follow an exponential distribution with a rate of $\frac{\binom{k}{2}}{N_y}$, $N_y \in \{17000, 23000\}$. This was verified by coalescence interval sizes recorded into 20 bins, and comparing the numbers in each bin to those expected if they were distributed $Exp(\frac{\binom{k}{2}}{N_y})$, using a chi-square test with 19 degrees of freedom.

Validation of Species Divergence Time Move

Given a null alignment, fixed population sizes in all edges of N_y, and a fixed gene tree with two leaves (one from each population), we would expect a well-behaved MCMC chain to return samples in which the difference between the root divergence time and the gene tree root to be exponentially distributed with a rate of $\frac{1}{N_y}$.

This is because the species root split depth is bounded on the lower end by 0, and on the upper end by the depth of the gene tree root. The species root divergence time is free to vary between these two limits, and the difference between the root divergence time and the upper limit is the effective time allowed for coalescence between the uncoalesced lineages from either population.

We implemented this test using population sizes of 1000. A total of 10 replicates were carried out, with the MCMC chain run for 1,000,000 steps and sampled every 10,000 steps. The samples of the root divergence time from the posterior were binned into 20 bins, and the proportion in each were compared to that expected under a exponential distribution with a rate of $\frac{1}{1000}$ using a chi-square test with 19 degrees of freedom.

Validation of Population Size Move

Given a null alignment, a fixed divergence time of 0, and a fixed gene tree with the coalescent intervals on the tree given by the expected values under the coalescent with a population size of N_y, we would expect a well-behaved MCMC chain to return a posterior sample of population sizes with a 95% HPD (high posterior density interval) that includes N_y.

This test was implemented using population sizes of $N_y \in \{17000, 49000\}$, with the MCMC chain run for 1,000,000 steps and sampled every 10,000 steps, for a total of 5 replicates per test condition.

Validation of Mutation Rate Move

Given a fixed divergence time of 0, a fixed population size of 30,000, a gene tree simulated under the coalescent with a population size of 30,000, and an alignment simulated on the gene tree with a mutation rate of μ_0, we would expect a well behaved MCMC chain to return a posterior sample of population sizes with a 95% HPD (high posterior density interval) that includes μ_0.

This test was implemented using mutation rates of $\mu_0 = 1e-6$, with the MCMC chain run for 1,000,000 steps and sampled every 10,000 steps, for a total of 5 replicates per test condition.

Validation of Divergence Time Model Jumping Move

Given a simple two-species system, with a single gene tree for each species, and a single gene sampled from each population of each species, we can analytically solve for the marginal likelihoods for the merged, single, divergence time model as well as the split, multiple divergence time model.

Let T_1 be the divergence time of an ancestral population, and g_1 be the time to the most recent common ancestor (MRCA) for a pair of genes sampled from each of the daughter populations. Let T_2 be the divergence time of a second ancestral population, and g_2 be the time to the most recent common ancestor (MRCA) for a pair of genes sampled from each of its daughter populations. Let $\mathcal{M}_1$ be the shared divergence time model, where $T_1 = T_2 = T_s$. Let $\mathcal{M}_2$ be the multiple divergence time model, where T_1 and T_2 are independent. Let the prior on divergence times be $\sim Uniform(0, u)$.

Shared Divergence Time Assuming without loss of generality that $g_1 \leq g_2$, the marginal likelihood of $\mathcal{M}_1$ is given by:

$$\begin{aligned}
\Pr(\mathcal{M}_1) &= \int_{T_0=0}^{g_1} K(g_1 - T_0)K(g_2 - T_0)\Pr(T_0)\, dT_0 \\
&= \int_{T_0=0}^{g_1} \frac{1}{N_1} e^{-\frac{1}{N_1}(g_1 - T_0)} \frac{1}{N_2} e^{-\frac{1}{N_2}(g_2 - T_0)} \frac{1}{u}\, dT_0 \\
&= \frac{1}{N_1 N_2 u} \int_{T_0=0}^{g_1} e^{-\frac{1}{N_1}(g_1 - T_0) - \frac{1}{N_2}(g_2 - T_0)}\, dT_0 \\
&= \frac{1}{N_1 N_2 u} \int_{T_0=0}^{g_1} e^{\frac{(N_1+N_2)T_0 - N_2 g_1 - N_1 g_2}{N_1 N_2}}\, dT_0. \qquad (3.22)
\end{aligned}$$

Let $x = \frac{(N_1+N_2)T_0 - N_2 g_1 - N_1 g_2}{N_1 N_2}$.

Then,

$$\begin{aligned}
dx &= \frac{N_1 + N_2}{N_1 N_2} dT_0 \\
dT_0 &= \frac{N_1 N_2}{N_1 + N_2} dx. \qquad (3.23)
\end{aligned}$$

Substituting expression 3.23 into 3.22, we get:

$$\frac{1}{N_1 N_2 u} \int_{x_0}^{x_1} \frac{N_1 N_2}{N_1 + N_2} e^x dx, \qquad (3.24)$$

where x_0 is given by:

$$x_0 = \frac{-N_2 g_1 - N_1 g_2}{N_1 N_2}, \qquad (3.25)$$

and x_1 is given by:

$$x_1 = \frac{g_1 - g_2}{N_2}. \qquad (3.26)$$

Solving the integral in expression 3.24, and substituting in the new limits:

$$\begin{aligned}\Pr(\mathcal{M}_1) &= \frac{1}{(N_1+N_2)u} e^x \Bigg|_{\frac{-N_2 g_1 - N_1 g_2}{N_1 N_2}}^{\frac{g_1-g_2}{N_2}} \\ &= \frac{1}{(N_1+N_2)u}\left[e^{\frac{g_1-g_2}{N_2}} - e^{\frac{-N_2 g_1 - N_1 g_2}{N_1 N_2}}\right] \end{aligned} \tag{3.27}$$

Multiple Divergence Times Assuming without loss of generality that $g_1 \leq g_2$, the marginal likelihood of $\mathcal{M}_2$ is given by:

$$\begin{aligned}\Pr(\mathcal{M}_2) &= \int_{T_1=0}^{g_1}\int_{T_2=0}^{g_2} K(g_1-T_1)\frac{1}{u}K(g_2-T_2)\frac{1}{u}\,dT_2 dT_1 \\ &= \int_{T_1=0}^{g_1}\int_{T_2=0}^{g_2} \frac{1}{N_1}e^{-\frac{1}{N_1}(g_1-T_1)}\frac{1}{u}\frac{1}{N_2}e^{-\frac{2}{N_2}(g_2-T_2)}\frac{1}{u}\,dT_2 dT_1 \\ &= \frac{1}{N_1 N_2 u^2}\int_{T_1=0}^{g_1}\int_{T_2=0}^{g_2} e^{-\frac{1}{N_1}(g_1-T_1)-\frac{1}{N_2}(g_2-T_2)}\,dT_2 dT_1 \\ &= \frac{1}{N_1 N_2 u^2}\int_{T_1=0}^{g_1}\int_{T_2=0}^{g_2} e^{\frac{T_1}{N_1}-\frac{g_1}{N_1}+\frac{T_2}{N_2}-\frac{g_2}{N_2}}\,dT_2 dT_1. \end{aligned} \tag{3.28}$$

Solving this integral yields:

$$\Pr(\mathcal{M}_2) = \frac{1}{u^2}\left(1 - e^{-\frac{g_1}{N_1}} - e^{-\frac{g_2}{N_2}} + e^{-\frac{g_1}{N_1}-\frac{g_2}{N_2}}\right). \tag{3.29}$$

From equations 3.27 and 3.29, for any particular value of g_1, g_2, N_1, N_2, and $\frac{1}{u}$, the marginal likelihoods of $\mathcal{M}_1$ vs. $\mathcal{M}_2$ can be evaluated, and from this the Bayes factor for $\mathcal{M}_1$ vs. $\mathcal{M}_2$, K, can be assessed directly. By setting $g_1 = 1e6$, $g_2 = 1391206$, $N_1 = N_2 = 1e5$, and $u = 1e7$, the Bayes factor K, comes very close to 1 (i.e., both models are equally favored, with $\frac{\Pr(\mathcal{M}_1|\mathbf{X})}{\Pr(\mathcal{M}_2|\mathbf{X})} = 1.000009$). Running an MCMC chain on this system should therefore yield the shared divergence time model being sampled as frequently as the multiple divergence time model.

`beluga` was run on a dataset corresponding to this system 1000 independent times. A χ^2 test

was carried out on each independent run, with

$$\chi^2 = \frac{(|\mathcal{M}_1| - 500)^2}{500} + \frac{(|\mathcal{M}_2| - 500)^2}{500}. \tag{3.30}$$

This statistic was compared to a χ^2 distribution with 1-degree of freedom.

3.4.2 Results

Validation of Gene Subtree Sliding Move

Each test replicate involved 19 independent chi-square tests (one for each coalescence event in a 20-leaf tree), resulting in a total of 380 chi-square tests (19 tests per replicate, for 10 replicates for each of the two population sizes). A sample plot of the results for one of the replicates is shown in Figure 3.1, showing the predicted and observed value in each bin. In 363 of these 380 tests, the null hypothesess could not be rejected at a 95% significance level. In the remaining tests, the null hypotheses were rejected with p-values ranging from 0.01 to 0.04. These latter cases are consistent with Type I error rate under the 95% significance level.

Validation of Species Divergence Time Move

Of the total of 10 test replicates carried out, eight could not be rejected at a 95% significance level. The null hypotheses were rejected in the remaining two tests with p-values of 0.04789 and 0.0377. A sample plot of the results for one of the replicates is shown in Figure 3.2.

Validation of Population Size Move

Figure 3.3 shows the combined densities for the posterior of the population size across all test replicates. The mean estimated population size across all runs was 16,9995.55.

Validation of Mutation Rate Move

Figure 3.4 shows the combined densities for the posterior of the mutation rate across all test replicates. In each test replicate, the true mutation rate ($1e^{-6}$) was within the 95% high-posterior density (HPD) region. The mean estimated mutation rate across all runs was 0.0000009697.

Validation of Divergence Time Model Jumping Move

In 934 of the 1000 independent tests (93.4% of the cases), the null hypothesis of equal probabilities of $\mathcal{M}_1$ vs. $\mathcal{M}_2$ could not be rejected at the 95% significance level. This is somewhat higher than the expected number of false rejections, and the discrepency could be attributed to numerical stability, statistical power, or, indeed, and actual bias in favor of the single divergence time model. More work is needed to address this issue.

3.5 Application to Simulated Data

3.5.1 Methods

The previous chapter discussed a two-species system that was used to evaluate an Approximate Bayesian Computation (ABC) approach to simultaneous divergence time estimation, `msbayes` (Hickerson et al., 2006a; Huang et al., 2011). Here, we apply `beluga` to the same data sets under two estimation conditions: one in which the true genealogies are given and one in which they need to be estimated along with the rest of the parameters.

Each data set consisted of a pair of species, and each species consisted of a pair of populations that diverged T_* generations in the past, $T_* \in \{0N, 4N, 8N, 32N, 64N\}$, $N = 2500$. In true simultaneous divergence, $T_1 = T_2 = 0$, while in the remaining cases the differences in divergence times ranged from $4N$ to $64N$. Samples were take from each population are various intervals after the second divergence event $(4N, 8N, 16N, 24N, 32N, 64N)$, with each set of samples from from each sampling period from each replicate of the experiment analyzed separately.

Genealogies were simulated using `Ginkgo` (Sukumaran and Holder, 2011), a forward-time spatially-explicit phylogeographic simulator. Three classes of simulations were carried out: a "baseline" class, in which all coalescent and model assumptions were met or approximated, a "migration" class in which there was post-vicariance gene flow of varying degrees, and and a "within-population structuring" class in which movement within a population was restricted to varying degrees. Sequences were simulated on the genealogies using `Seq-Gen` (Rambaut and Grassly, 1997) under two mutation rates, a per site per generation mutation rates of $2.4e^{-7}$ and $7e^{-6}$, for two categories of population parameter values, repsectively: a "low theta regime" ($\theta = 0.0024$) and a "high regime" ($\theta = 0.07$).

A total of 10 simulation replicates was generated for each distinct combination of classes, configurations, parameters and settings.

A further set of simulations was carried out using multi-locus data. Parameter sweeps were sparser in this case, focussing on a smaller subset of divergence time separations $(0N, 8N, 16N)$ and sampling periods $(4N, 16N, 32N)$.

The estimation conditions in which the true genealogies were given and fixed were used to assess the accuracy of the model selection under conditions of low-dimensionality / high information. Specifically, `beluga` was given the true gene trees along with the data, and these were fixed (i.e., the MCMC chain did not propose any changes to them). These analysis conditions will provide indications of the accuracy and performance of the more novel aspects of the `beluga` estimation procedure, namely the simultaneous divergence time model jumping, as opposed to the Bayesian inference of phylogenies, which is a well-established field (Felsenstein, 2004). Priors on population sizes were $\sim Uniform(0, 10000)$, while priors on the clade divergence times were $\sim Uniform(10000, 1000000)$. Each MCMC chain was run for 5,000,000 steps which took approximately 4 minutes to run to completion on a 3.3 GHz Intel Xeon machine. The MCMC chain was sampled every 5000 steps, for a total of 1000 samples from the posterior. Convergence was assessed by visual inspection of log-posterior, log-likelihood and parameter estimate densities of randomly selected runs using `Tracer` (Rambaut A., 2007). In all such inspected runs, convergence occurred very rapidly, and so the first 100 samples were conservatively discarded as burn-in across all runs. The posterior probability of the divergence time model (one, i.e., simultaneous divergence, or two, i.e., non-simultaneous or multiple independent divergence) was given directly by the proportional representation of the respective model in samples from the posterior.

The estimation procedure was also run with the unfixed gene trees, i.e, `beluga` had to estimate the gene trees and mutation rate along with all the other parameters under a low theta regime. All priors were the same as before, with the additional mutation rate prior given as $\sim Uniform(1e^{-8}, 1e^{-6})$. Each MCMC chain was run for 250,000,000 steps which took approximately 12 hours to run to completion on a 3.3 GHz Intel Xeon machine. The MCMC chain was sampled every 250,000 steps, for a total of 1000 samples from the posterior. Again, covergence was assessed by visual inspection of log-posterior, log-likelihood and parameter estimate densities of randomly selected runs using `Tracer` (Rambaut A., 2007), and the first 100 samples were

discarded as burn-in.

3.5.2 Results

The results of the low-dimensionality (fixed, true gene trees) analyses of data generated under baseline conditions (all model assumptions met) under both low and high theta regimes are shown in Figures 3.5 and 3.6, respectively. Non-simultaneous divergence generally cannot be identified if the time separating two non-simultaneous events is $4N$ or less. At time separations of $8N$ or greater, however, non-simultaneous divergence can generally be identified correctly.

The results of the low-dimensionality analyses of data generated under conditions where the Wright-Fisher population assumptions are violated by the introduction of gene flow between the daughter sub-populations are shown in Figures 3.7 through 3.9. As more and more migration is introduced, there is a tendency to prefer a non-simultaneous divergence time model unconditionally, regardless of the actual generating model or truth.

The results of the low-dimensionality analyses of data generated under conditions where the Wright-Fisher population assumptions are violated by the introduction of substructuring within the daughter sub-populations are shown in Figures 3.10 through 3.12. As more and more sub-structuring is introduced, there is a tendency to prefer a non-simultaneous divergence time model unconditionally, regardless of the actual generating model or truth.

The results of the high-dimensionality analyses (gene trees are estimated as part of the inference) of data generated under baseline conditions (all model assumptions met) under low theta regimes are shown in Figure 3.14. Under this very parameter-rich model, performance is extremely poor, with more than $64N$ generations separating the two divergence events required before the non-simultaneous divergence model is correctly preferred.

3.6 Discussion

`beluga` provides a way to estimate simultaneously the divergence time between sister populations of multiple species in parallel, using data from multple loci and integrating information from coalescent, population genetic as well as phylogenetic processes in a Bayesian statistical framework. In its most general case, `beluga` can analyze an arbitrary number of independent species. However, if

limited to two species, the current version allows for reverse-jump MCMC to sample from models of different dimensionality with respect to the divergence time, so as to estimate explicitly the posterior probability of simultaneous divergence vs. non-simultaneous divergence. In this respect, its application domain is similar to that of `msbayes` (Huang et al., 2011; Hickerson et al., 2006b), and it can thus be seen as a full- or exact-likelihood counterpart. Tests and verification of various components of the model and the program in isolation all indicate that both the model and its implementation produce reasonable results as predicted by coalescent and Bayesian theory.

In applications to simulated data when given the true gene trees as information, performance was excellent as long as the assumptions of the model were not violated. In particular, under baseline conditions where coalescent assumptions of Wright-Fisher demographic condtions as well as complete post-vicariance isolation as assumed by the estimation model were met, `beluga` was able to identify correctly the number of distinct divergence times with a resoluton of $4N$ generations. That is, as long as $4N$ generations or more separated separated two independent divergence events, given the gene trees, `beluga` was consistently able to prefer correctly the non-simultaneous divergence model with very strong support. If the time separating the two divergences was less than $4N$ generations, on the other, `beluga` concluded preference for the non-simultaneous divergence model.

Migration or incomplete post-vicariance gene flow between the daughter populations misled the `beluga` analyses, as it did with the `msbayes` analyses that did not treat migration. However, the response was in the opposite direction, with spurious, strong, and unconditional support for the simultaneous divergence. Since migration is excluded from the `beluga` model, a migrant allele can only be explained by reducing the divergence time between daughter population such that the allele is modelled as a coalescent event in the ancestor. With sufficient levels of migration, the divergence times in both species get reduced to the extremely recent past, and thus any information regarding time separation is lost. In effect, untreated migration "blocks" the method from "seeing" any deeper in the past than the most recent migration, and a single unstructured population becomes generally a better explanation for data from both species.

Within population substructuring confounded or confused `beluga` in much the same way, both quantitatively and qualitatively, as it did `msbayes` and `IMa2`. As within-population substructuring levels increased, a general trend toward unconditional support for non-simultaneous divergence was observed, even if the divergence was actually simultaneous. This suggests that there might be no

summary statistics that might rescue the Approximate Bayesian Computation approach from being misled by the within-population substructuring. The likelihood can be seen as a perfect summary statistic (Beaumont et al., 2010), and so if it lacks the power to discriminate correctly between simultaneous and non-simultaneous divergence in a coalescent framework, it is unlikely that this situation can be improved given the additional loss of information with any other coalescent-based summary statistic. Indeed, these analyses were provided more information than ever might be available to any real-world empirical study. The fixed gene trees alone reduced the dimensionality of the problem by several orders of magnitude: the only numerical parameters to be estimated were six population size parameters and two divergence time parameters. The resulting model is extremely simple, and hence highly tractable *and* powerful, as indicated by the fact that, without within-population structuring, `beluga` converged on the truth extremely rapidly and consistently across multiple tests. However, with strong within-population structuring, `beluga` failed every time. All this suggests that the limitation lies with the fact that if there substantial or strong population structuring, then the coalescent submodel of any estimation framework needs to explicitly account or treat this.

The high-dimensionality analyses produce poor results. When comparing these to `msbayes` or `IMa2`, it should be remembered that the `beluga` analyses were under models of much higher dimensionality than either these previous two. The `msbayes` model reports its results in units of N, and thus implicitly assumes equal population sizes. The `IMa2` model, as used in the previous chapter, explicitly assumes an equal mutation rate, and when comparing analyses across independent runs, equal population sizes were also assumed. In the high-dimensionality `beluga` analyses presented, all mutation rates and population sizes were free parameters, for a total of eight extra free parameters. Furthermore, some of these free parameters are known to be non-identifiable given the existing free parameters (e.g., population size, divergence time, gene tree edge lengths, mutation rate). Given all this, it is actually surprising that `beluga` was able to identify correctly the divergence time model at all. The fact that it could do this at the most extreme time separation analyzed indicates the power of coalescent-based statistical phylogeography. At the same time, it also speaks to the limits of knowledge: without some information on at least one if not more classes of the parameters (mutation rate, population size, divergence time), then phylogeographic model selection in general, and simultaneous divergence time testing in particular, is extremely difficult if not impossible.

A comprehensive sensitivity analysis of `beluga` is needed, to establish the full boundaries of its performance envelope. There are strong inter-dependencies between many of the parameters of the `beluga` model (e.g., population size and divergence time, mutation rate and divergence time), and informative priors on some of these may allow for better power in estimating others.

`beluga` is currently limited in a number of ways, all of which have the potential to be improved in future work. The most obvious limitation is the reverse-jump MCMC between divergence time schedule models is limited to two species. There is no theoretical reason that this cannot be extended to an arbitary number of species, and this extension will greatly increase the applicability and the usefulness of this program.

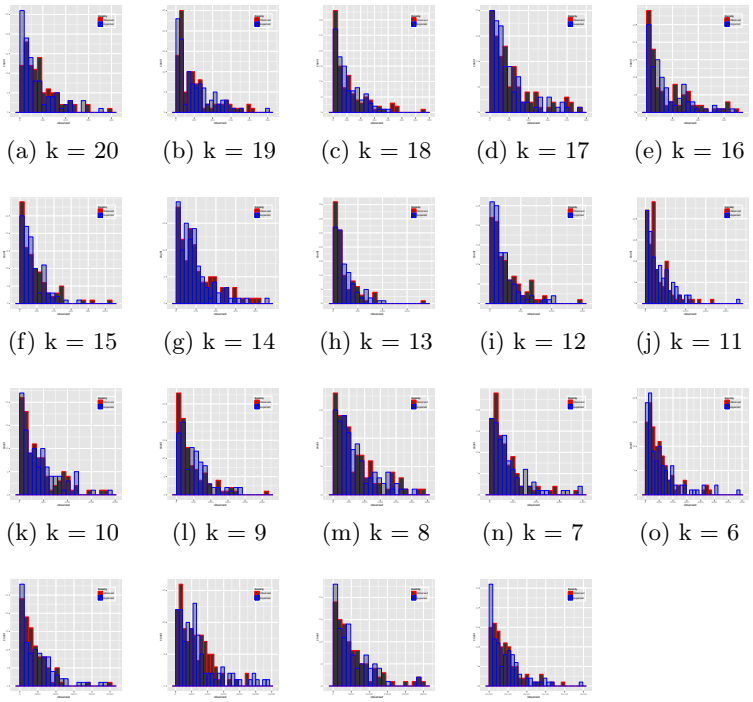
Furthermore, population relationships are currently limited to a single split (i.e., each species tree is assumed to only have two leaves). This situtation is also easily resolved, and will be one of the targets for future work.

Another area for improvement is the efficiency of some of the MCMC moves. In particular, there are a number of sophisticated algorithms for ultrametric or clock-constrained phylogenetic tree proposals (Höhna et al., 2008), some of which may perform much better than the simple subtree slide move used here, and others of which might work well in conjunction with it. Improved efficiency or performance in searching through gene tree space will directly contribute to better usage of multiple locus information. In addition, the model jumping currently uses a very crude scheme to propose the divergence time of a species when splitting away (i.e., transitioning from the single divergence time state to a multiple divergence time state). This leads an extremely high proportion of rejections of this move, which slows down mixing considerably. A more flexible move may propose a new divergence time for the split group that, for example, decays exponentially from the current divergence time, to increase the probability of it being accepted.

If the parameter space has multiple peaks, and, given the complexity of the space as well as the well known non-identifiability of certain groups of parameters in some of the submodels, (such as the mutation rate and time, or population size and time) then Metropolis coupling (Geyer, 1991; Altekar et al., 2004) or will increase performance of the MCMC chains. Finally, while within-population structuring present a challenging prospect to be treated statisically, models incorporating post-vicariance migration have been developed and implemented (Hey and Nielsen, 2004, 2007). Future work on `beluga` will incorporate these aspects, and allow for application in a broader range of

contexts.

3.7 Figures



(a) k = 20 (b) k = 19 (c) k = 18 (d) k = 17 (e) k = 16

(f) k = 15 (g) k = 14 (h) k = 13 (i) k = 12 (j) k = 11

(k) k = 10 (l) k = 9 (m) k = 8 (n) k = 7 (o) k = 6

(p) k = 5 (q) k = 4 (r) k = 3 (s) k = 2

Figure 3.1: Results of a single replicat of the gene subtree slide move validation test. Each subplot, (a) through (s), shows the observed (red) vs. expected (blue) frequencies of coalescent interval sizes. The expected distribution should follow a coalescent distribution for a population size of 17000 and k lineages, $Exp(\frac{\binom{k}{2}}{17000}), k \in \{2, 3, ..., 20\}$. See text for details.

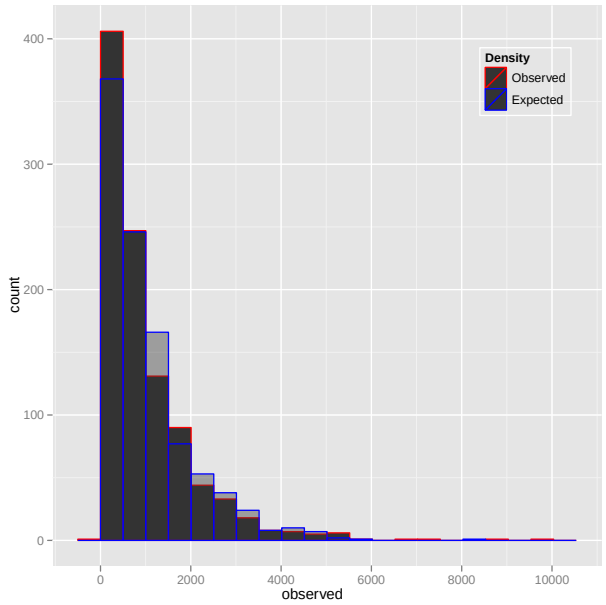


Figure 3.2: Results of a single replicate of the divergence time move validation test, showing observed (red) vs. expected (blue) frequencies of differences in divergence time vs. root of genealogy for a population size of 17000. The expected distribution should follow a coalescent with 2 lineages and a haploid population size of 17000, i.e. $Exp(\frac{1}{17000})$.

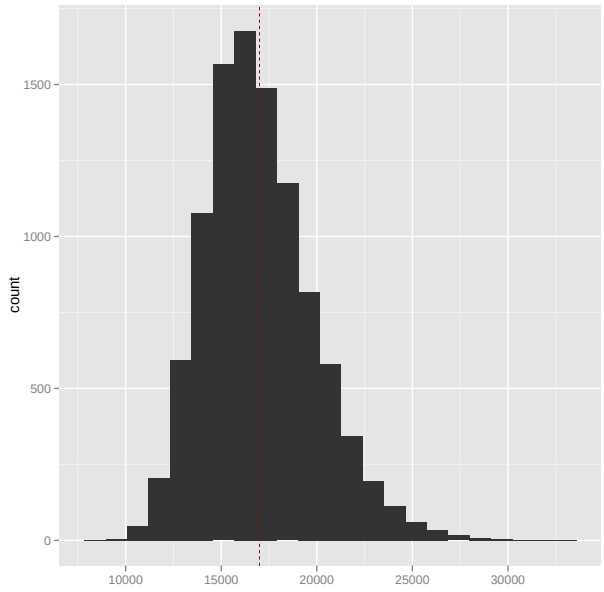


Figure 3.3: Results of combined results of 10 independent tests of the population size MCMC move/estimation procedure, showing estimated posterior distribution of across all runs. True population size was 17000 (indicated by red dotted line). Mean of posterior was 16995.5478164, while the 95% HPD was $(11956.958467587323, 22799.177855160102)$.

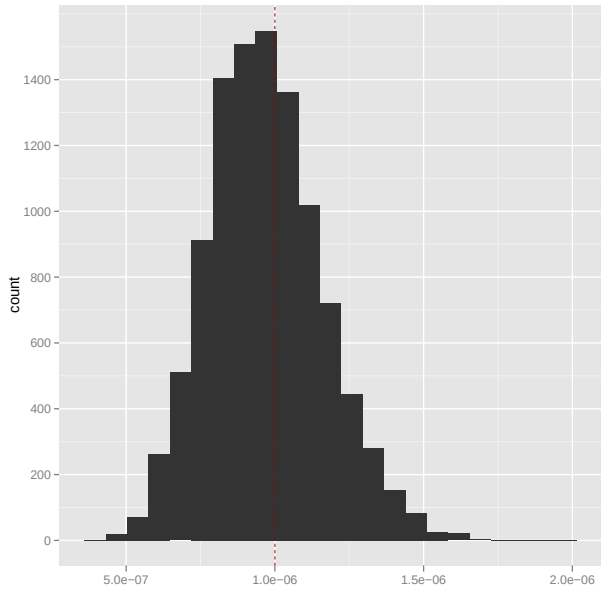


Figure 3.4: Results of combined results of 10 independent test of the mutation rate MCMC move/estimation procedure, showing estimated posterior distribution of across all runs. True mutation rate was $1e-6$ (indicated by red dotted line). Mean of posterior was 9.69e-7, while the 95% HPD was $(6.17805e-07, 1.357387e-06)$.

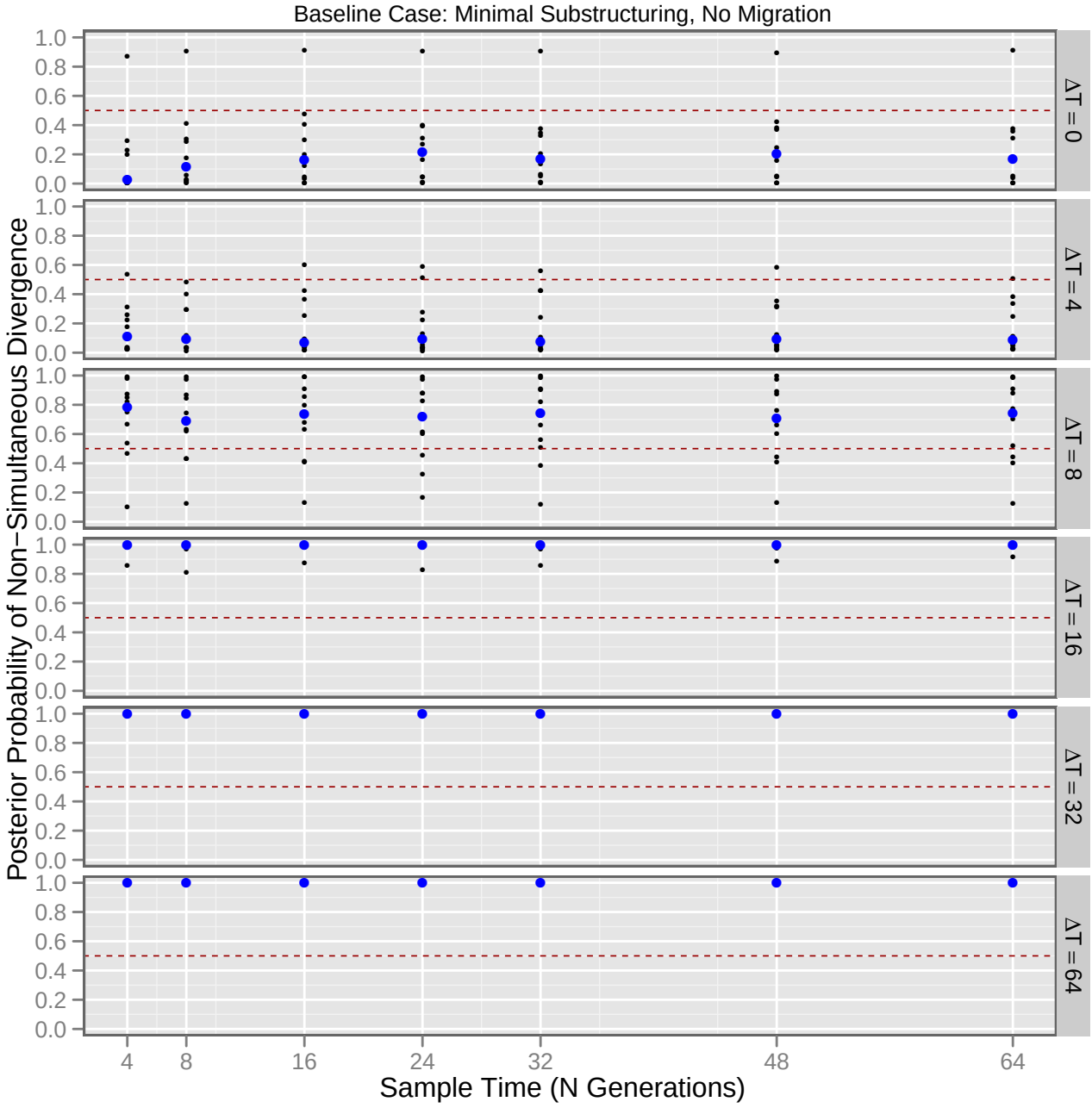


Figure 3.5: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under *low* theta values when analyzed using `beluga`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. Y-axis on each strip indicates estimated approximate posterior probability of support for *multiple* divergences, while X-axis indicates period after the second vicariance event in which the sample was taken. Each strip is for 10 replicates of a simulation carried out at particular difference or separation of time between the two divergences. The top strip, $\Delta T = 0$, is when there is *no* difference in time between the two divergences, i.e., the case of true simultaneous divergence. High values on the Y-axis here indicate support for the *wrong* model. The remaining strips show increasingly larger differences in time between divergence events, and high values on the Y-axis thus indicate support for the *correct* model.

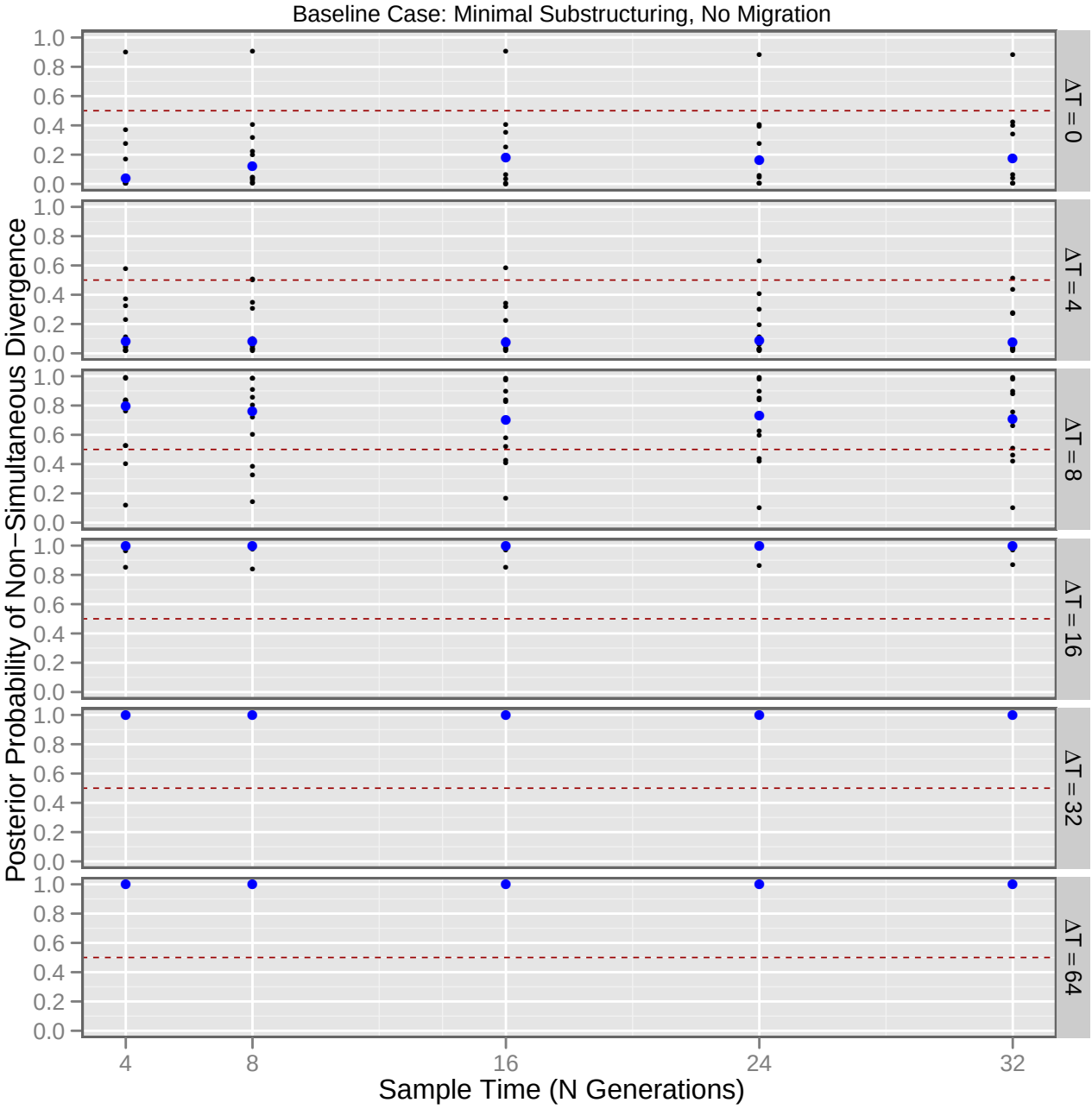


Figure 3.6: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under *high* theta values when analyzed using `beluga`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 3.5 for details on interpreting the plots.

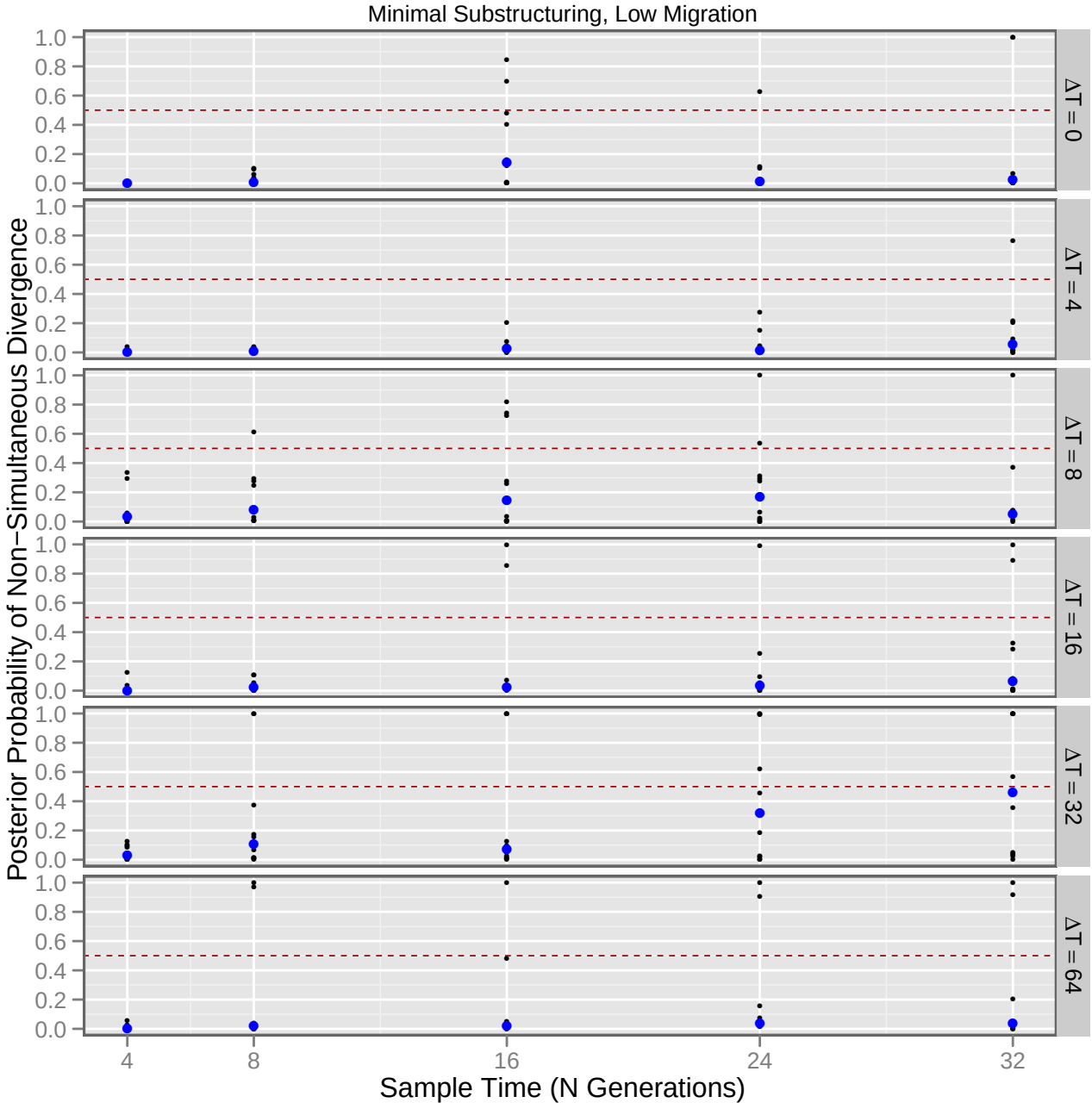


Figure 3.7: Posterior probability of non-simultaneous divergence of forward-time simulations under *high* theta values and *low* post-vicariance gene flow when analyzed using `beluga`. See figure 3.5 for details on interpreting the plots.

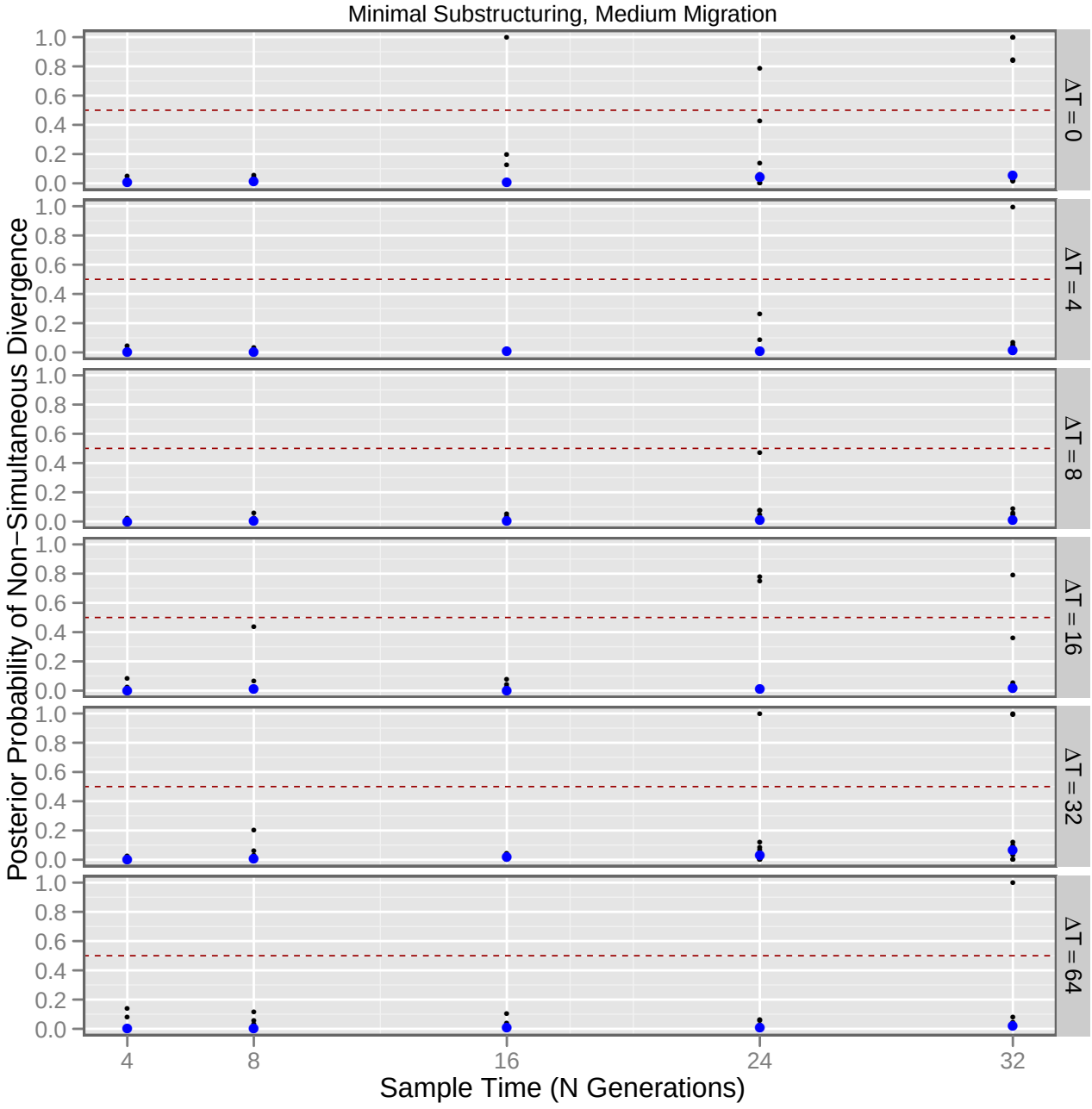


Figure 3.8: Posterior probability of non-simultaneous divergence of forward-time simulations under *high* theta values and *medium* post-vicariance gene flow when analyzed using `beluga`. See figure 3.5 for details on interpreting the plots.

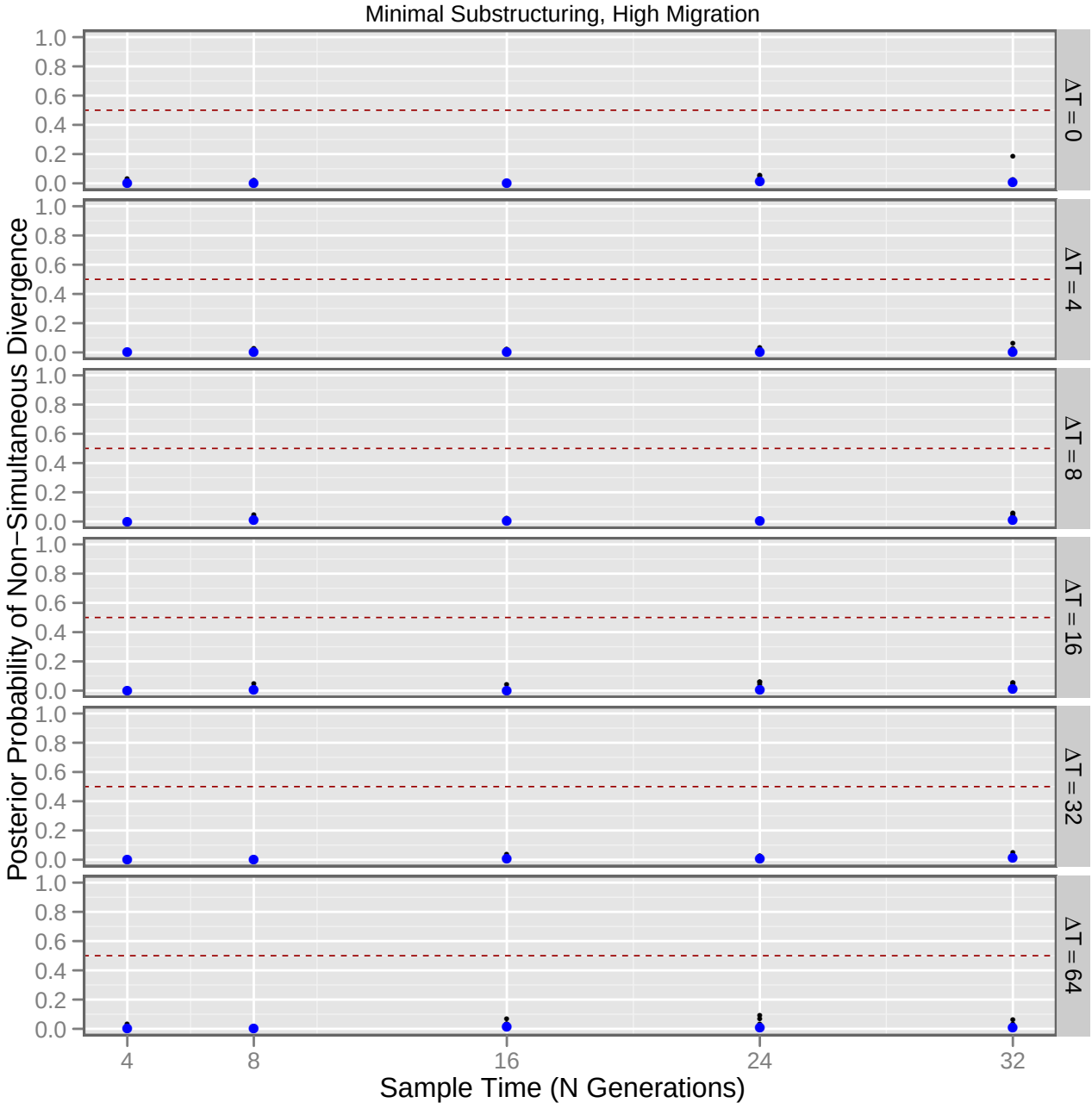


Figure 3.9: Posterior probability of non-simultaneous divergence of forward-time simulations under *high* theta values and *high* post-vicariance gene flow when analyzed using `beluga`. See figure 3.5 for details on interpreting the plots.

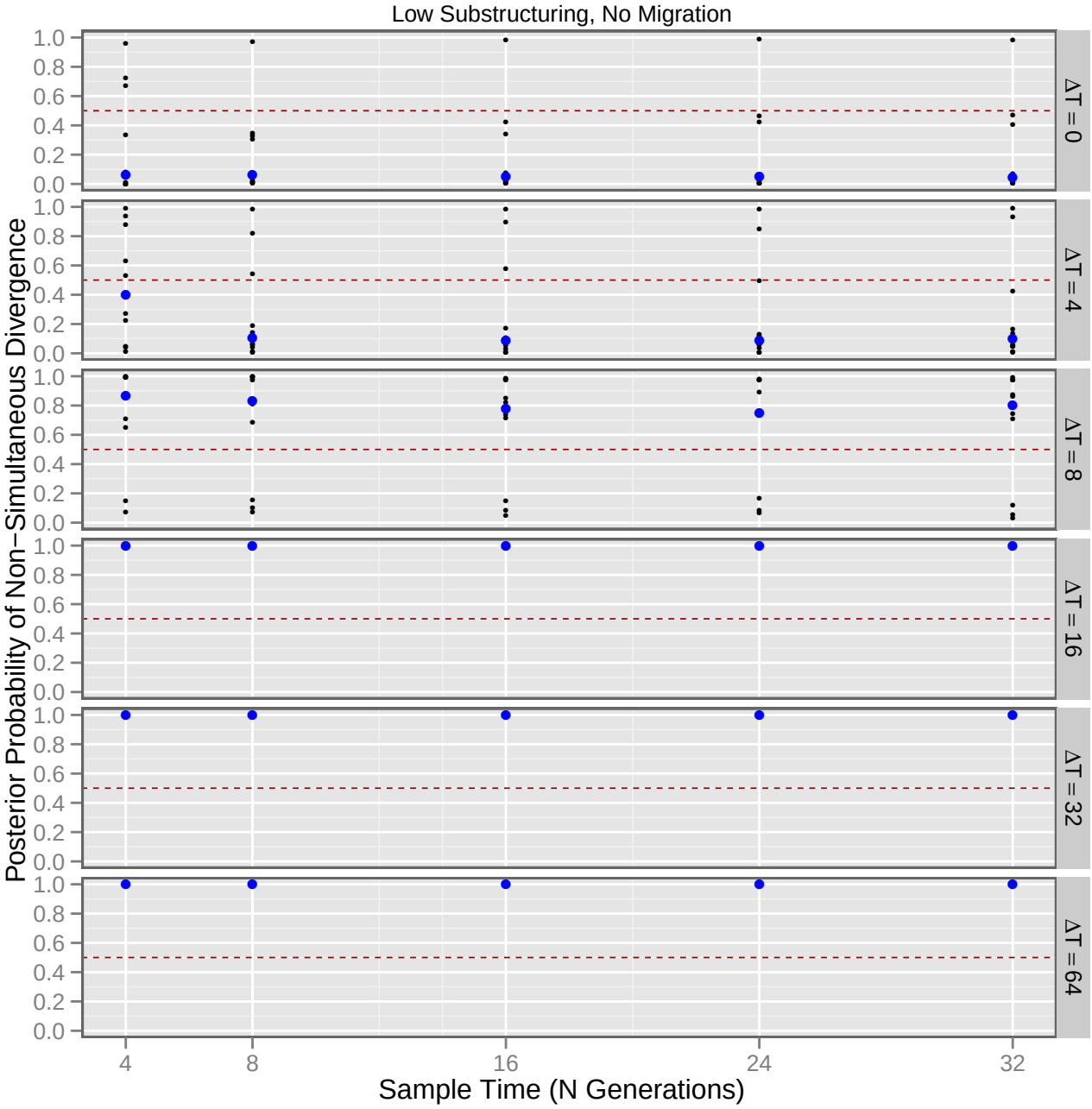


Figure 3.10: Posterior probability of non-simultaneous divergence of forward-time simulations under *high* theta values and *low* within-population substructuring when analyzed using `beluga`: the Wright-Fisher assumptions of the estimation model were selectively violated by introducing a low degree of movement restriction within daughter subpopulations. See figure 3.5 for details on interpreting the plots.

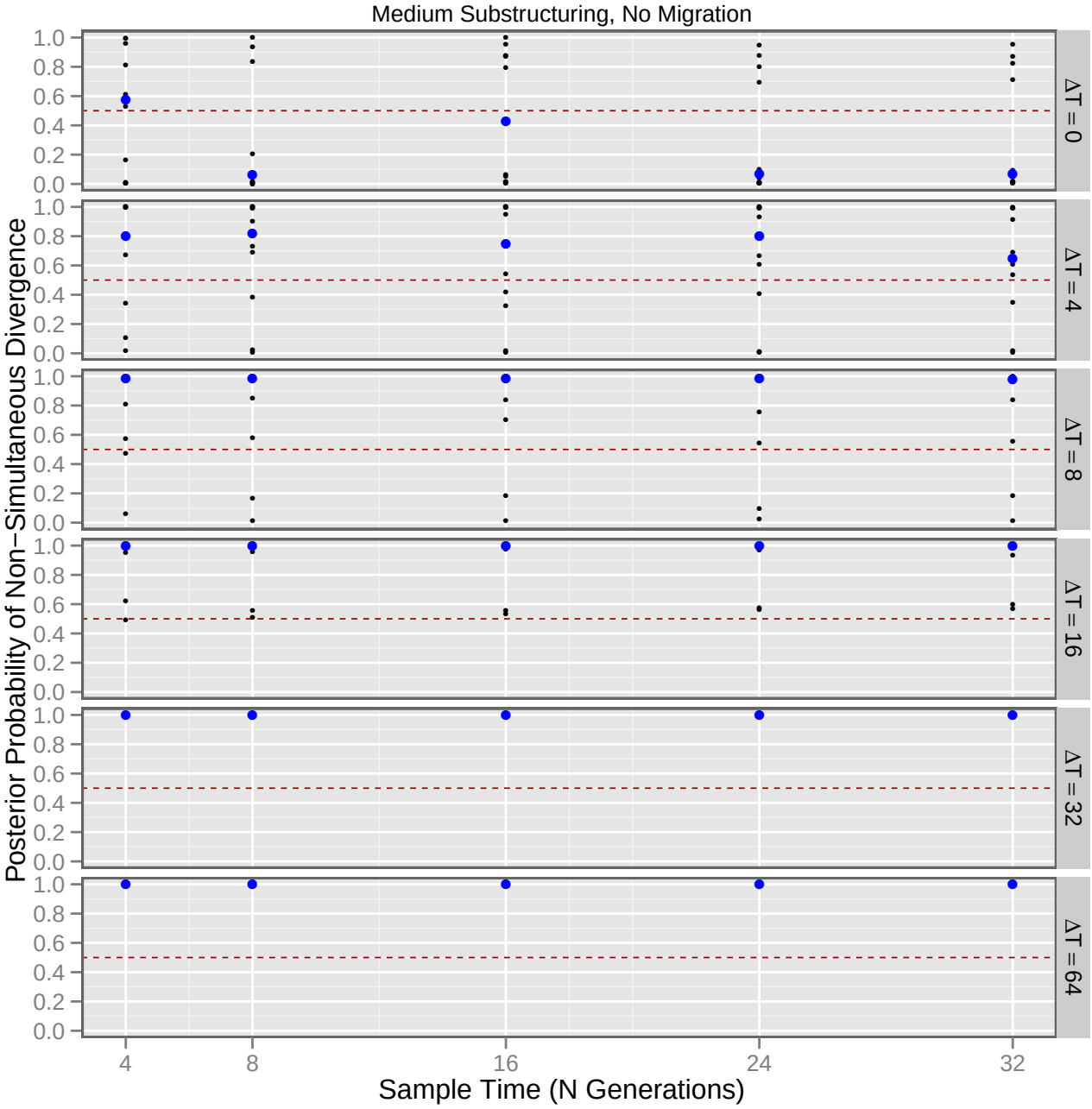


Figure 3.11: Posterior probability of non-simultaneous divergence of forward-time simulations under *high* theta values and *medium* within-population substructuring when analyzed using `beluga`: the Wright-Fisher assumptions of the estimation model were selectively violated by introducing a medium degree of movement restrictions within daughter subpopulations. See figure 3.5 for details on interpreting the plots.

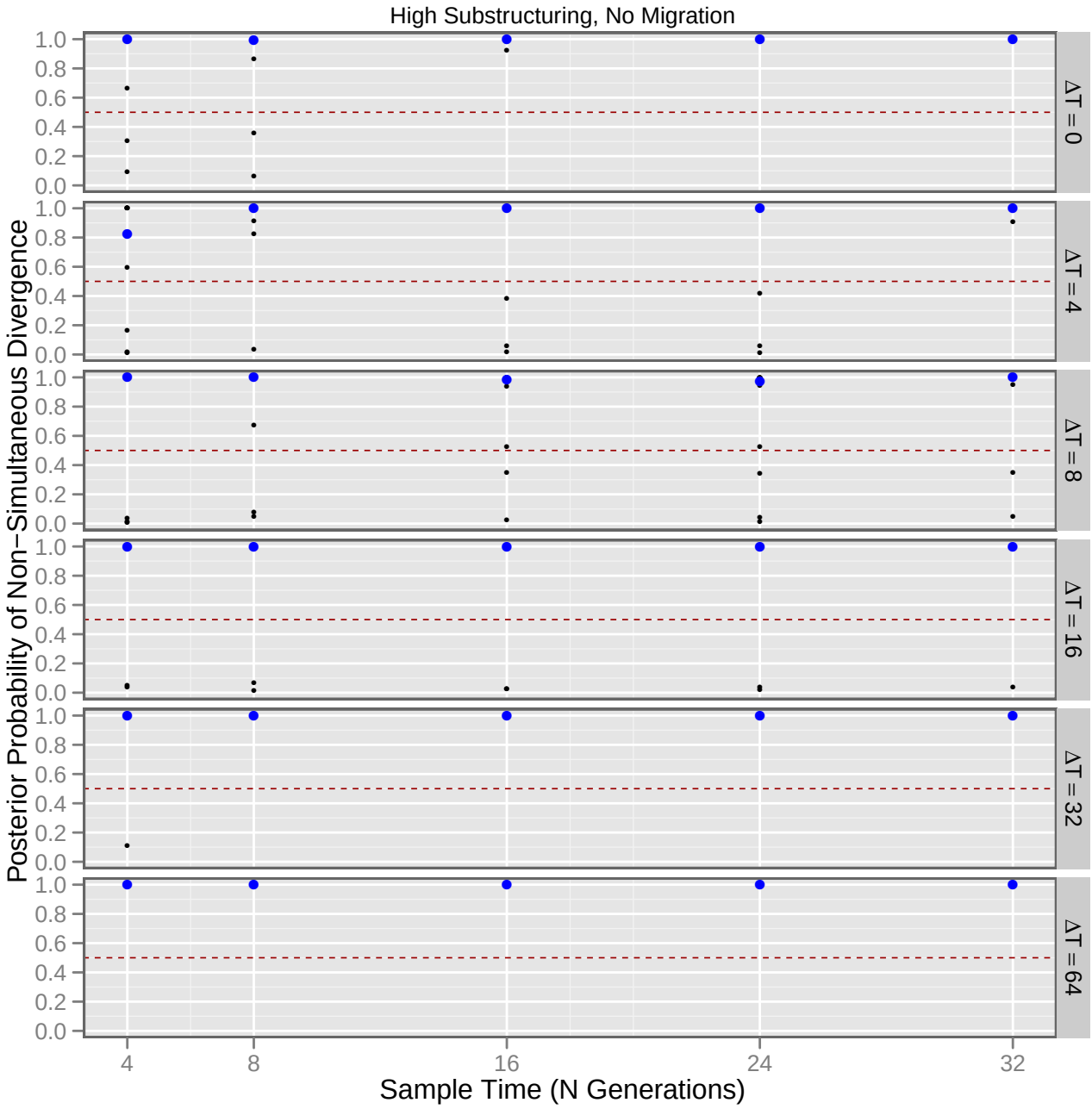


Figure 3.12: Posterior probability of non-simultaneous divergence of forward-time simulations under *high* theta values and *high* within-population substructuring when analyzed using `beluga`: the Wright-Fisher assumptions of the estimation model were selectively violated by introducing a high degree of movement restrictions within daugher subpopulations. See figure 3.5 for details on interpreting the plots.

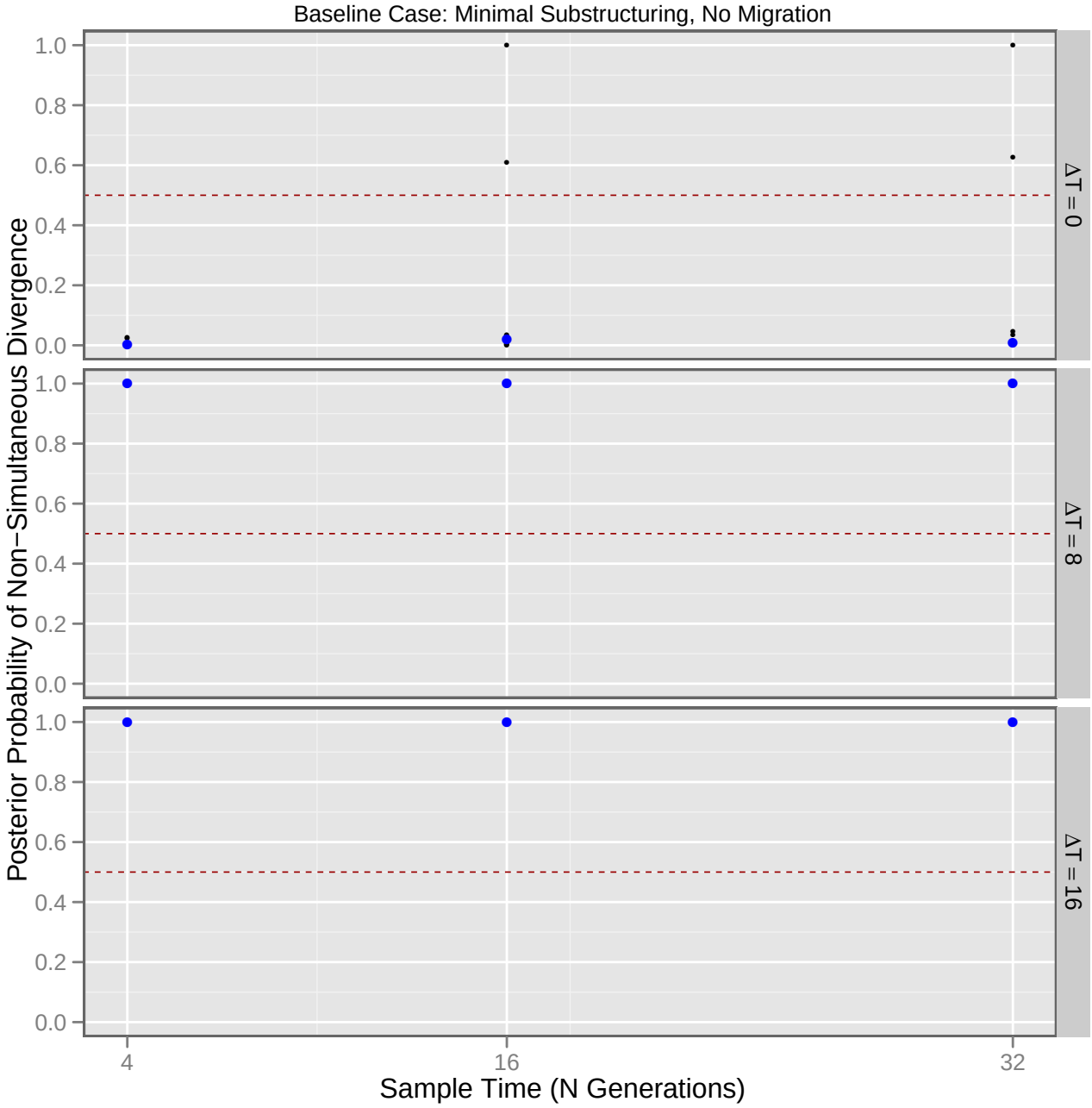


Figure 3.13: Posterior probability of non-simultaneous divergence of multilocus (5 loci) "baseline" forward-time simulations under *low* theta values when analyzed using `beluga`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. See figure 3.5 for details on interpreting the plots.

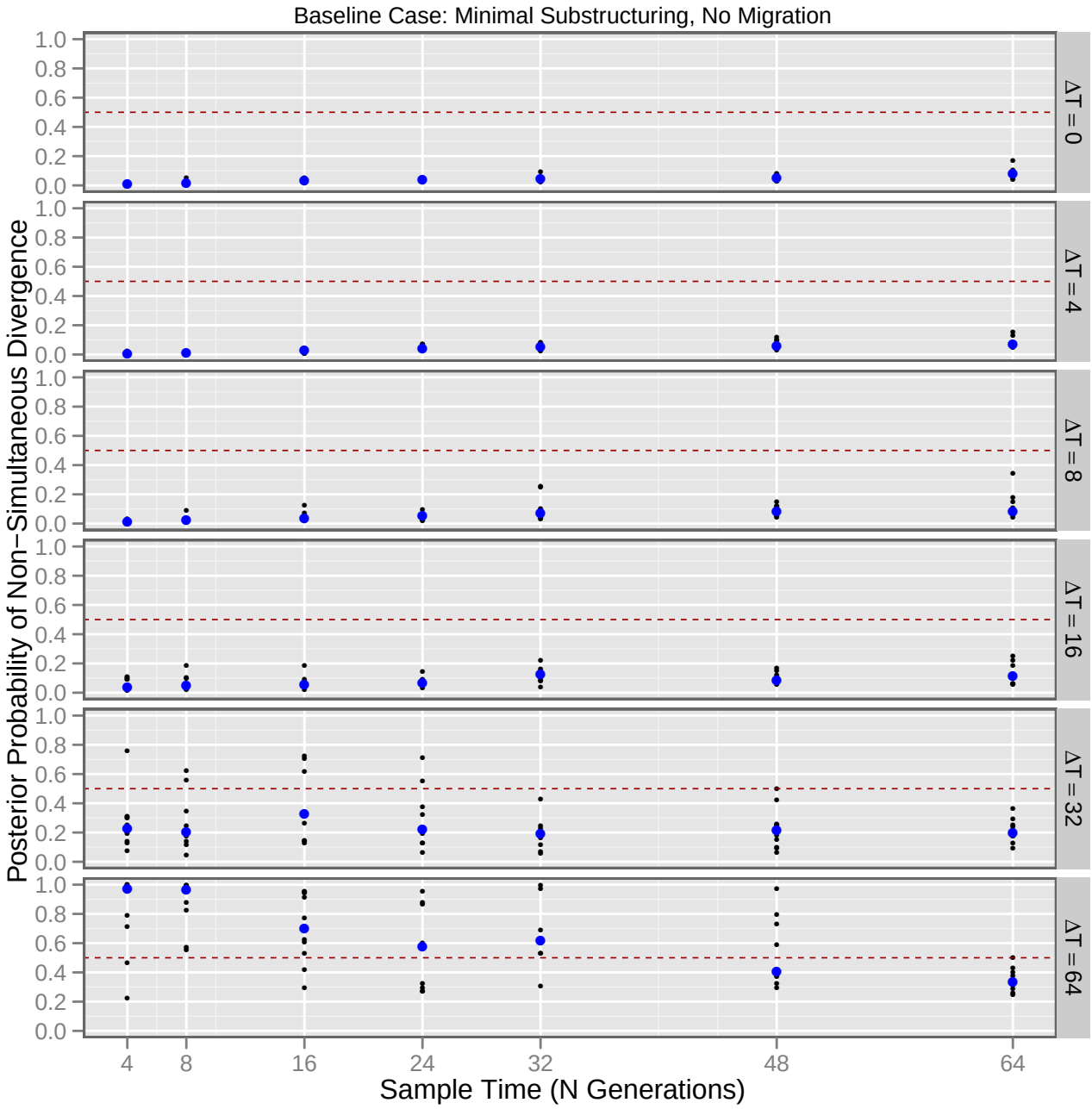


Figure 3.14: Posterior probability of non-simultaneous divergence of "baseline" forward-time simulations under *low* theta values when analyzed using `beluga`: all assumptions of the estimation model, in particular, Wright-Fisher population assumptions and complete post-vicariance isolation (no migration) were met. MCMC was over gene tree parameter space as well. See figure 3.5 for details on interpreting the plots.

Part V

Appendices

References

Bibliography

Altekar, G., S. Dwarkadas, J. Huelsenbeck, and F. Ronquist, 2004. Parallel metropolis coupled markov chain monte carlo for bayesian phylogenetic inference. Bioinformatics 20:407–415.

Avise, J., 2000. Phylogeography: the history and formation of species. Harvard University Press.

Ayres, D., A. Darling, D. Zwickl, P. Beerli, M. Holder, P. Lewis, J. Huelsenbeck, F. Ronquist, D. Swofford, M. Cummings, et al., 2011. Beagle: an application programming interface and high-performance computing library for statistical phylogenetics. Systematic Biology .

Baker, W., M. Coode, J. Dransfield, and S. Dransfield, 1998. Patterns of distribution of Malesian vascular plants. Biogeography and Geological Evolution of SE Asia .

Barber, B. R. and J. Klicka, 2010. Two pulses of diversification across the Isthmus of Tehuantepec in a montane Mexican bird fauna. Proceedings Of The Royal Society B-Biological Sciences 277:2675–2681.

Beaumont, M., 2010. Approximate bayesian computation in evolution and ecology. Annual Review of Ecology, Evolution, and Systematics 41:379–406.

Beaumont, M., R. Nielsen, C. Robert, J. Hey, O. Gaggiotti, L. Knowles, A. Estoup, M. Panchal, J. Corander, M. Hickerson, et al., 2010. In defence of model-based inference in phylogeography. Molecular Ecology 19:436–446.

Beaumont, M., W. Zhang, and D. Balding, 2002. Approximate bayesian computation in population genetics. Genetics 162:2025–2035.

Beerli, P. and J. Felsenstein, 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 of the United States of America 98:4563.

Bertorelle, G., A. Benazzo, and S. Mona, 2010. ABC as a flexible framework to estimate demography over space and time: some cons, many pros. Molecular Ecology 19:2609–2625.

Brown, J., K. Savidge, and E. McTavish, 2009. Demography and Individual Migration Simulated Using a Markov chain. Molecular Ecology Resources 11:2623–2635.

Brown, R. M. and S. I. Guttman, 2002. Phylogenetic systematics of the rana signata complex of philippine and bornean stream frogs: reconsideration of huxley's modification of wallace's line at the oriental- australian faunal zone interface. Biological Journal of the Linnean Society .

Carling, M. and R. Brumfield, 2007. Gene sampling strategies for multi-locus population estimates of genetic diversity (θ). PLoS One 2:e160.

Carstens, B., J. Degenhardt, A. Stevenson, and J. Sullivan, 2005. Accounting for coalescent stochasticity in testing phylogeographical hypotheses: modelling Pleistocene population structure in the Idaho giant salamander *Dicamptodon aterrimus*. Molecular Ecology 14:255–265.

Carstens, B., A. Stevenson, J. Degenhardt, and J. Sullivan, 2004. Testing nested phylogenetic and phylogeographic hypotheses in the *Plethodon vandykei* species group. Systematic Biology 53:781–792.

Currat, M., N. Ray, and L. Excoffier, 2004. 2004. SPLATCHE: a program to simulate genetic diversity taking into account environmental heterogeneity 4:139–142.

Daza, J. M., T. A. Castoe, and C. L. Parkinson, 2010. Using regional comparative phylogeographic data from snake lineages to infer historical processes in Middle America. Ecography 33:343–354.

Didelot, X., R. Everitt, A. Johansen, and D. Lawson, 2011. Likelihood-free estimation of model evidence. Bayesian Analysis 6:49–76.

Evans, B., R. Brown, J. McGuire, and J. Supriatna, 2003. Phylogenetics of fanged frogs: Testing biogeographical hypotheses at the interface of the Asian and Australian faunal zones. Systematic Biology 52:794–819.

Excoffier, L., 1995. Amova 1.55 (analysis of molecular variance). Genetics and Biometry Laboratory, University of Geneva, Geneva, Switzerland .

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

———, 2004. Inferring Phylogenies. 1 ed. Sinauer Associates, Inc, Sunderland, Massachusetts.

Fisher, R. A., 1930. The Genetical Theory of Natural Selection. Oxford Univeristy Press.

Garrick, R., R. Dyer, L. Beheregaray, and P. Sunnucks, 2008. Babies and bathwater: a comment on the premature obituary for nested clade phylogeographical analysis. Molecular Ecology 17:1401–1403.

Geyer, C., 1991. Markov chain Monte Carlo maximum likelihood. Defense Technical Information Center.

Green, P., 2003. Trans-dimensional markov chain monte carlo. Oxford Statistical Science Series Pp. 179–198.

Green, P. J., 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Hall, R., 2001. Cenozoic reconstructions of SE Asia and the SW Pacific: changing patterns of land and sea. Faunal and Floral Migrations and Evolution in Se Asia- ... .

Hasegawa, M., H. Kishino, and T. Yano, 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. Journal of Molecular Evolution 22:160–174. URL http://www.springerlink.com/index/G22V36XLM59W6826.pdf.

Hastings, W. K., 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Heaney, L., J. Walsh Jr, and A. Townsend Peterson, 2005. The roles of geological history and colonization abilities in genetic differentiation between mammalian populations in the philippine archipelago. Journal of Biogeography 32:229–247.

Hey, J., 2006. Recent advances in assessing gene flow between diverging populations and species. Current opinion in genetics & development 16:592–596.

———, 2010. Isolation with migration models for more than two populations. Molecular Biology and Evolution 27:905–920.

Hey, J. and R. Nielsen, 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. Genetics 167:747–760.

———, 2007. Integration within the felsenstein equation for improved markov chain monte carlo methods in population genetics. Proceedings of the National Academy of Sciences 104:2785.

Hickerson, M. J., G. Dolman, and C. Moritz, 2006a. Comparative phylogeographic summary statistics for testing simultaneous vicariance. Molecular Ecology 15:209–223.

Hickerson, M. J., E. Stahl, and H. Lessios, 2006b. Test for simultaneous divergence using approximate Bayesian computation. Evolution 60:2435–2453.

Hickerson, M. J., E. Stahl, and N. Takebayashi, 2007. msBayes: Pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. BMC Bioinformatics 8:268.

Höhna, S., M. Defoin-Platel, and A. Drummond, 2008. Clock-constrained tree proposal operators in bayesian phylogenetic inference. *in* BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on, Pp. 1–7. IEEE.

Holder, M., 2011. Python wrapper around beagle-lib library for calculation of likelihoods on phylogenetic trees. https://github.com/mtholder/pytbeaglehon.

Holloway, J. and R. Hall, 1998. SE Asian geology and biogeography: an introduction. Biogeography and Geological Evolution of SE Asia .

How, R. and D. Kitchener, 1997. Biogeography of Indonesian snakes. Journal of Biogeography .

Huang, W., N. Takebayashi, Y. Qi, and M. Hickerson, 2011. Mtml-msbayes: Approximate bayesian comparative phylogeographic inference from multiple taxa and multiple loci with rate heterogeneity. BMC Bioinformatics 12:1. URL http://www.biomedcentral.com/1471-2105/12/1.

Hudson, R., 2002. Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Huizinga, D. and A. Kolawa, 2007. Automated defect prevention: best practices in software management. Wiley-IEEE Computer Society Pr.

Inger, R., 1999. Distribution patterns of amphibians in Souther Asia and adjacent islands. Pp. 1–30, *in* W. Duellman, ed. Patterns of distribution of amphibians: a global perspective. Johns Hopkins Univ Pr.

Inger, R. F., 2005. The frog fauna of the Indo-Malayan region as it applies to Wallace's line. Wallace in Sarawak–150 Years Later. An International Conference on Biogeography and Biodiversity Pp. 82–89.

Irwin, D., 2002. Phylogeographic breaks without geographic barriers to gene flow. Evolution 56:2383–2394.

Kingman, J., 1982a. The coalescent. Stochastic Processes and their Applications 13:235–248.

———, 1982b. On the genealogy of large populations. Journal of Applied Probability 19:27–43.

Knowles, L. L., 2008. Why does a method that fails continue to be used? Evolution 62:2713–2717.

Knowles, L. L. and W. P. Maddison, 2002. Statistical phylogeography. Molecular Ecology 11:2623–2635.

Kuhner, M., P. Beerli, J. Yamato, and J. Felsenstein, 2000. Usefulness of single nucleotide polymorphism data for estimating population parameters. Genetics 156:439.

Leache, A. D., S. C. Crews, and M. J. Hickerson, 2007. Two waves of diversification in mammals and reptiles of Baja California revealed by hierarchical Bayesian analysis. Biology Letters 3:646–650.

Leuenberger, C. and D. Wegmann, 2010. Bayesian computation and model selection without likelihoods. Genetics 184:243–52.

Lewis, P., 2003. Ncl: a c++ class library for interpreting data files in nexus format. Bioinformatics 19:2330–2331.

Lourie, S. and A. Vincent, 2004. A marine fish follows Wallace's Line: the phylogeography of the three-spot seahorse (Hippocampus trimaculatus) in Southeast Asia. Journal of Biogeography .

Maddison, D., D. Swofford, and W. Maddison, 1997. Nexus: an extensible file format for systematic information. Systematic Biology 46:590.

Marin, J.-M., P. Pudlo, C. Robert, and R. Ryder, 2011. Approximate bayesian computational methods. Statistics and Computing Pp. 1–14. URL http://dx.doi.org/10.1007/s11222-011-9288-2. 10.1007/s11222-011-9288-2.

Mayr, E., 1944. Wallace's line in the light of recent zoogeographic studies. The Quarterly Review of Biology 19:1–14.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1092.

Michaux, B., 2010. Biology of Wallacea: geotectonic models, areas of endemism, and natural biogeographic units. Biological Journal of the Linnean Society 101:193–212.

Morley, R. and J. Flenley, 1987. Late cainozoic vegetational and environmental changes in the malay archipelago. Biogeographical evolution of the Malay Archipelago 50:1259–1264.

Nei, M., 1975. Molecular population genetics and evolution P. 704.

Nei, M., A. Chakravarti, and Y. Tateno, 1977. Mean and variance of FST in a finite number of incompletely isolated populations. Theoretical Population Biology 11:291 – 306.

Nielsen, R., 2000. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics 154:931.

Nielsen, R. and M. Beaumont, 2009. Statistical inferences in phylogeography. Molecular ecology 18:1034–1047.

Panchal, M., 2007. The automation of nested clade phylogeographic analysis. Bioinformatics 23:509–510.

Panchal, M., M. Beaumont, and P. Sunnucks, 2007. The automation and evaluation of nested clade phylogeographic analysis. Evolution 61:1466–1480.

Petit, R., 2008. The coup de grâce for the nested clade phylogeographic analysis? Molecular Ecology 17:516–518.

Rambaut, A. and N. Grassly, 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Bioinformatics 13:235.

Rambaut A., D. A. J., 2007. Tracer v1.4. http://beast.bio.ed.ac.uk/Tracer.

Robert, C., J.-M. Marin, and N. S. Pillai, 2011. Why approximate Bayesian computational (ABC) methods cannot handle model choice problems URL http://arxiv.org/abs/1101.5091.

Schulte, J. A. I., J. Melville, and A. L. Larson, 2003. Molecular phylogenetic evidence for ancient divergence of lizard taxa on either side of Wallace's LINE. Proceedings of the Royal Society B: Biological Sciences .

Simpson, G., 1977. Too many lines; the limits of the oriental and australian zoogeographic regions. Proceedings of the American Philosophical Society 121:107–120.

Sukumaran, J. and M. Holder, 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

Sukumaran, J. and M. T. Holder, 2011. Ginkgo: spatially-explicit simulator of complex phylogeographic histories. Molecular Ecology Resources 11:364–369. URL http://dx.doi.org/10.1111/j.1755-0998.2010.02926.x.

Swofford, D., 1998. PAUP 4.0: phylogenetic analysis using parsimony. Smithsonian Institution.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by dna polymorphism. Genetics Pp. 585–595.

Templeton, A., 2008. Nested clade analysis: an extensively validated method for strong phylogeographic inference. Molecular Ecology 17:1877–1880.

Templeton, A. R., 2009. Why does method that fails continue to be used? The answer. Evolution 63:807–812.

Turner, H., P. Hovenkamp, and P. van Welzen, 2001. Biogeography of Southeast Asia and the West Pacific. Journal of Biogeography .

Van Oosterzee, P., 1997. Where worlds collide: the Wallace Line. Cornell Univ Pr.

Wakeley, J., 1996. Distinguishing migration from isolation using the variance of pairwise differences. Journal of Theoretical Population Biology 49:369–386.

———, 2009. Coalescent Theory, an Introduction. Roberts and Company, Greenwood Village, CO. URL http://www.coalescentheory.com/.

Watterson, G. A., 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256 – 276. URL http://www.sciencedirect.com/science/article/pii/0040580975900209.

Waxman, D., 2006. Fisher's geometrical model of evolutionary adaptation–beyond spherical geometry. Journal of Theoretical Biology 241:887–895.

Wegmann, D., C. Leuenberger, S. Neuenschwander, and L. Excoffier, 2010. Abctoolbox: a versatile toolkit for approximate bayesian computations. BMC Bioinformatics 11:116. URL http://www.biomedcentral.com/1471-2105/11/116.

Whitmore, T., 1987. Biogeographical evolution of the Malay Archipelago. Oxford.: Clarendon Press.

Woodruff, D. S., 2003. Neogene marine transgressions, palaeogeography and biogeographic transitions on the Thai-Malay Peninsula. Journal of Biogeography 30:551–467.

Yang, Z., 2006. Computational molecular evolution. Oxford University Press, USA.

Zhu, H., P. A. V. Hall, and J. H. R. May, 1997. Software unit test coverage and adequacy. ACM Computing Surveys 29:366–427.