Supplementary Information for:

Estimation of species divergence times in presence of cross-species gene flow

George P. Tiley[1†], Tomás Flouri[2], Xiyun Jiao[2,3], Jelmer W. Poelstra[1], Bo Xu[4], Tianqi Zhu[5,6], Bruce Rannala[7], Anne D. Yoder[1*], Ziheng Yang[2*]

[1]Department of Biology, Duke University, Durham, NC, USA

[2]Department of Genetics, Evolution and Environment, University College London, London, UK

[3]Department of Statistics and Data Science, China Southern University of Science and Technology, Shenzhen, Guangdong, China

[4]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

[5]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China

[6]Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, China

[7]Department of Evolution and Ecology, University of California, Davis, Davis, CA, USA

[†]Present address: Royal Botanic Gardens Kew, Richmond, Surrey, UK

Supplementary Methods

Parameter estimates from simulated data

Parameters were estimated from simulated data with BPP v4.1.4 for both the MSC and MSci models. We used a burnin of 20,000 MCMC iterations and collected 200,000 samples sampling every 2 iterations. The prior on θ was an inverse gamma with shape parameter $\alpha = 3$ and scale parameter $\beta = 2\theta$. This is a diffuse prior with the prior mean $\frac{\beta}{(\alpha-1)}$ at the true value. Similarly, the prior on the root age τ_{r} was an inverse gamma with $\alpha = 3$ and $\beta = 6\theta$ so that the prior mean is 3θ. The other node ages follow a uniform Dirichlet distribution (Yang and Rannala 2010, Eq. 2). For the MSci model only, a prior Beta distribution with both shape parameters set at 1 was specified on φ, which is equivalent to the uniform distribution for the interval from 0 to 1. Visual inspection of a few analyses from the simulated data were done to assure MCMC settings were appropriate, and that parameter estimates had converged.

Additional scrutiny was given to seemingly aberrant runs. Inspection of MCMC samples and posterior summaries revealed that some replicates converged to the true value with efficient MCMC sampling while some tended to samples with biologically implausibly large θ and wide HPD intervals despite the posterior median being very close to the true value. Were we to collect many more MCMC samples for these simulation analyses with poorly estimated θ from the MSC model, it is reasonable to expect improved estimates. However, we present these estimates as poor MCMC sampling occurs in empirical data analysis.

Calibrating divergence times

We considered two common approaches to calibrations that convert divergence times from substitutions per site to absolute time in years. First, we used a node calibration on the root, which could come from fossil evidence placed near the crown of a group (e.g. Benton and Donoghue 2007) or a well-justified geological event (e.g., Da Baets et al. 2016). Second, we used a substitution rate calibration. A de novo mutation rate could be used to calibrate divergences (Yoder and Tiley 2021), or a phylogenetic substitution rate from a previous study could be used as well (Campbell et al. 2021). It is up to the investigator to determine the best calibration, but we make the comparison here since a rate calibration will depend on the absolute node differences between models (as presented in our simulation results), while a node calibration depends only on the relative distances between nodes. The differences between calibration strategies could lead to different conclusions about absolute ages between the MSC and MSci models. For example a root node with an age of 0.01 substitutions per site under the MSC and 0.02 substitutions per site under the MSci would both have an age of 10 Ma ago if the same node calibration were applied. However, a rate calibration applied to all nodes of a tree would cause the MSci root divergence time to be twice as old as the MSC divergence time.

Here we apply simple node and rate calibrations, to demonstrate differences between the MSci and MSC model with real data rather than attempting to obtain the best possible absolute divergence times for the two systems. Since the rate calibration presented was based on substitutions per-site from the MSC analyses and we adhere to a strict molecular clock, the node-calibrated and rate-calibrated MSC analyses are equivalent.

Baobabs

We assumed that the sampled outgroup *Scleronema* diverged from *Adansonia* 18.2 Ma ago (Marinho et al. 2014). Divergence times were estimated under MSC and MSci using this node calibration by setting the mean posterior estimate of τ_c (Supplementary Figure S25) to 18.2 and rescaling all other mean estimates and HPD intervals by $\frac{node\ \mathrm{age}}{\tau_c} \times 18.2$. A rate calibration in substitutions per site per million years was obtained from the MSC model by $rate = \frac{\tau_c}{18.2} =$

$\frac{0.012883}{18.2} = 0.00070786$. The rate-calibrated MSci estimates were then generated by $\frac{node\ \text{age}}{rate}$. All MSC and MSci estimates in substitutions per site are available in Supplementary Table S1 and calculating the calibrated values is trivial given the assumption of a strict clock.

Jaltomata

The procedure for deriving node-calibrated and rate-calibrated divergence times was the same as for baobabs. As node calibration, we assumed that *Solanum* and *Jaltomata* diverged 17 Ma ago (τ_a in Supplementary Figure S32) (Sarkinen et al. 2013). All node ages under the MSC and MSci models in substitutions per site are in Supplementary Table S2.

References

Benton MJ, Donoghue PJC. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol.* 24:26-63.

Campbell CR, Tiley GP, Poelstra JW, Hunnicutt KE, Larsen PA, Lee H-J, Thorne JL, dos Reis M, Yoder AD. 2021. Pedigree-based and phylogenetic methods support surprising patterns of mutation rate and spectrum in the gray mouse lemur. *Heredity.* 127:23-244.

De Baets K, Antonelli A, Donoghue PJC. 2016. Tectonic blocks and molecular clocks. *Philos Trans R Soc Lond Ser B Biol Sci.* 371:20160098.

Marinho RC, Mendes-Rodrigues C, Balao F, Ortiz PL, Yamagishi-Costa J, Bonetti AM, Oliveira PE. 2014. Do chromosome numbers reflect phylogeny? New counts for Bombacoideae and a review of Malvaceae s.l. *Am J Bot.* 101:1456-1465.

Sarkinen T, Bohs L, Olmstead RG, Knapp S. 2013. A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. BMC Evol. Biol. 13:214.

Yang Z, Rannala B. 2010. Bayesian species delimitation using multilocus sequence data. *Proc Natl Acad Sci USA* 107:9264-9269.

Yoder AD, Tiley GP. 2021. The challenge and promise of estimating the de novo mutation rate from whole-genome comparisons among closely related individuals. *Mol Ecol.* 30:6087-6100.

Supplementary Figures

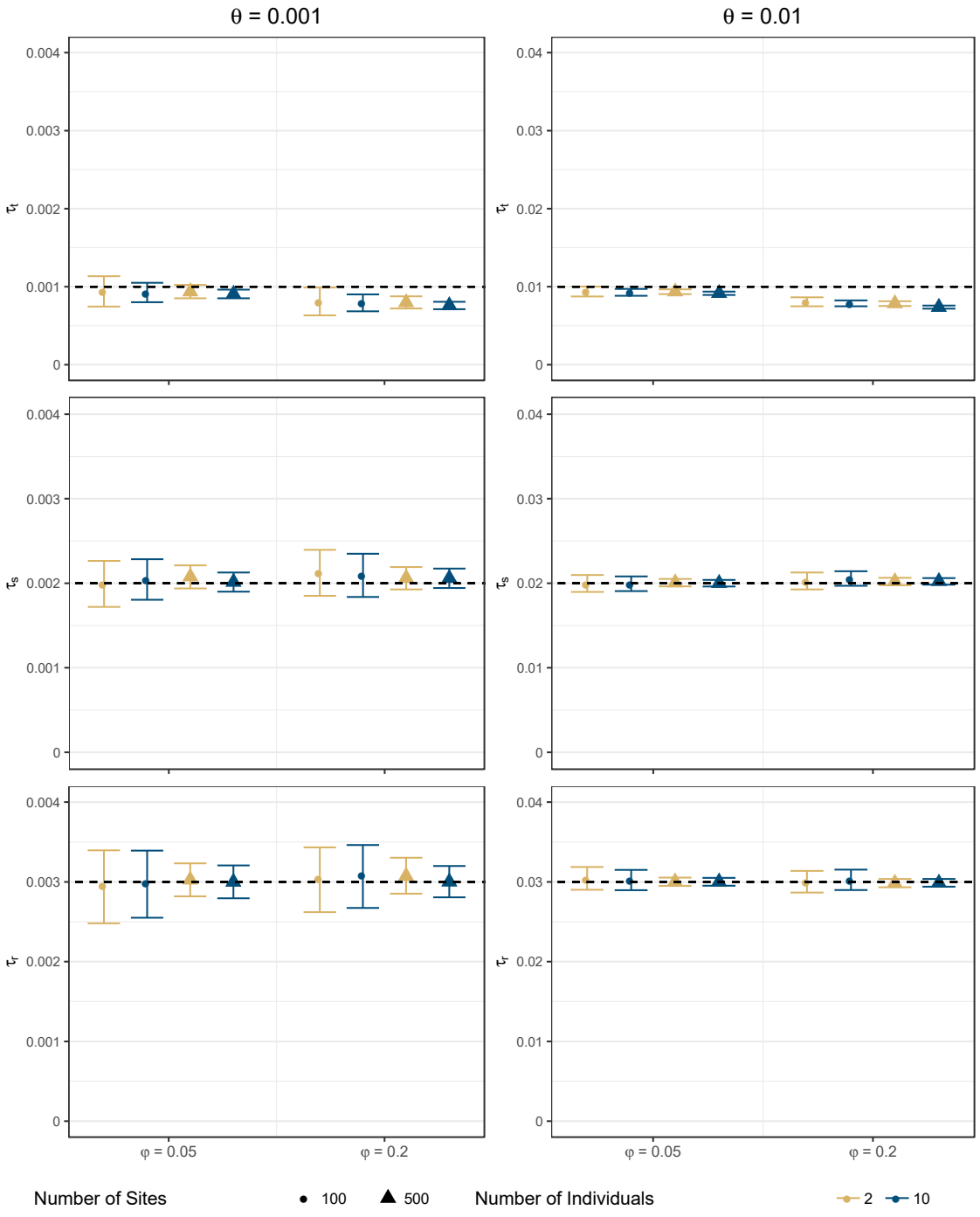

Figure S1 – Divergence Time Estimates under the MSC with Introgression Between Sister Lineages. Points are posterior means and bars are 95% HPD CIs. Dashed lines are true values. Node labels and divergence times correspond to Fig. 1a.

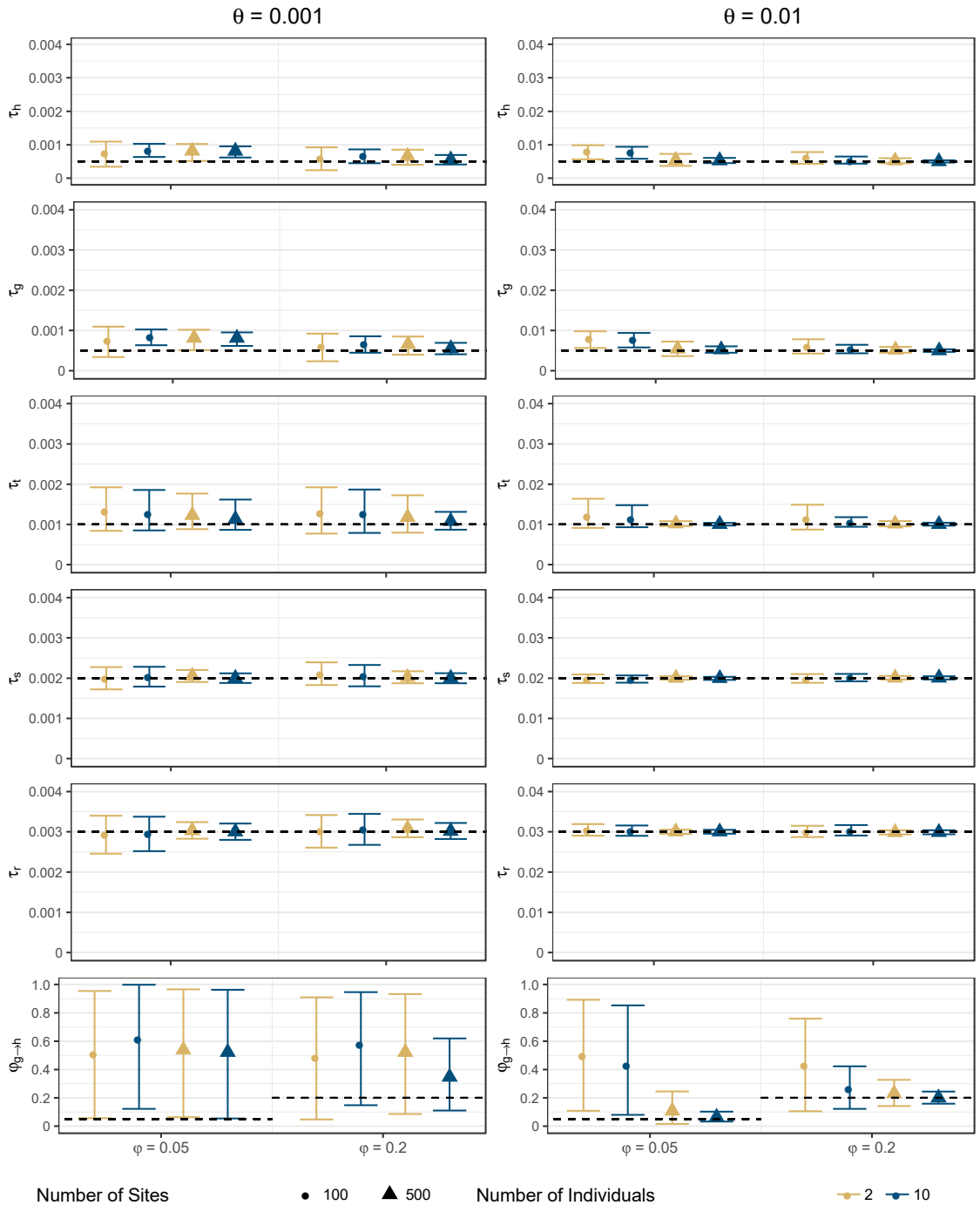

Figure S2 – Divergence Time and Introgression Probability Estimates Under the MSci with Introgression Between Sister Lineages. Points are posterior means and error bars are the 95% HPD CIs. Dashed lines are true values. Node labels and divergence times correspond to Fig. 1a.

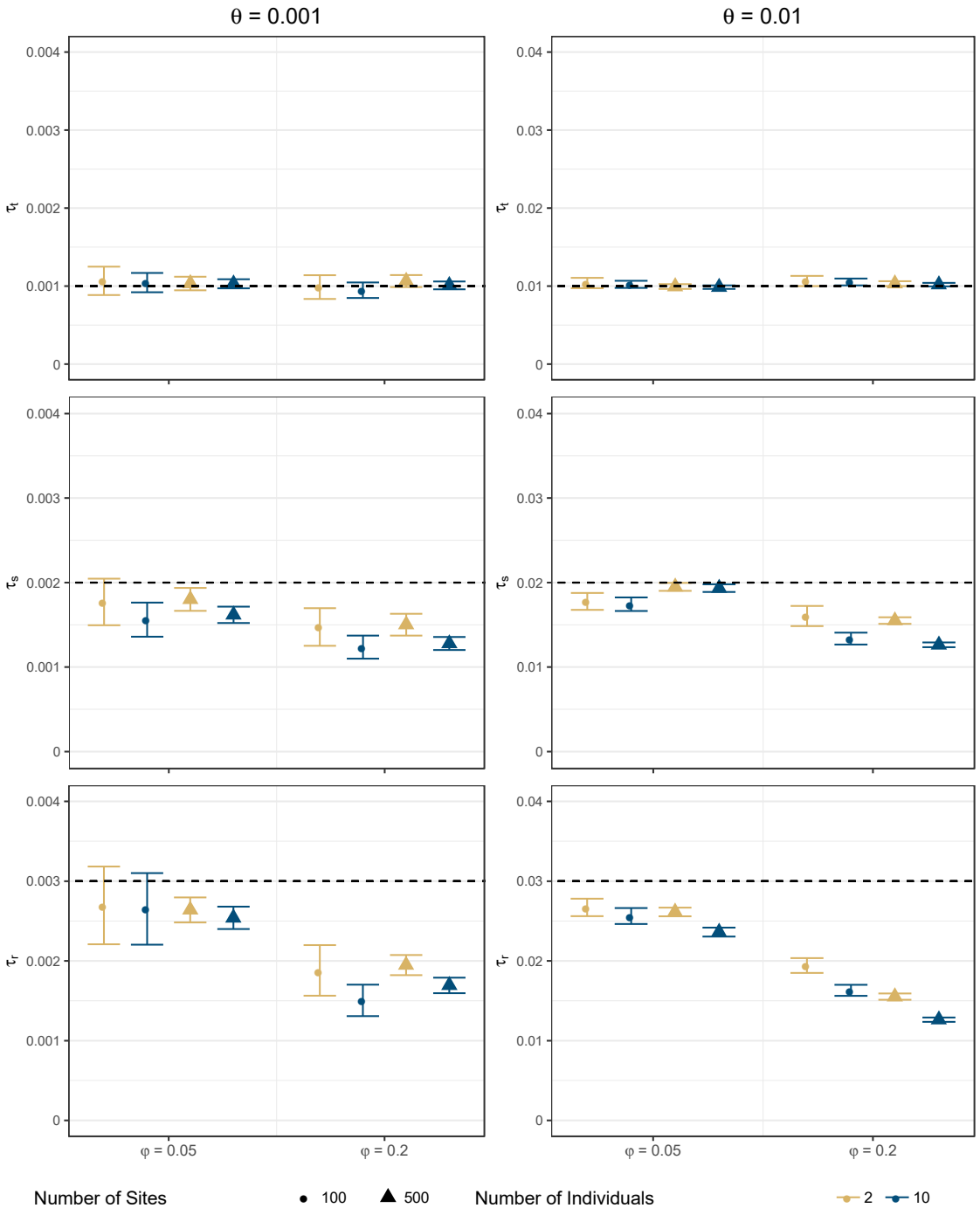

Figure S3 – Divergence Time Estimates under the MSC with Introgression Between Non-Sister Lineages. Node labels and divergence times correspond to Fig. 1b.

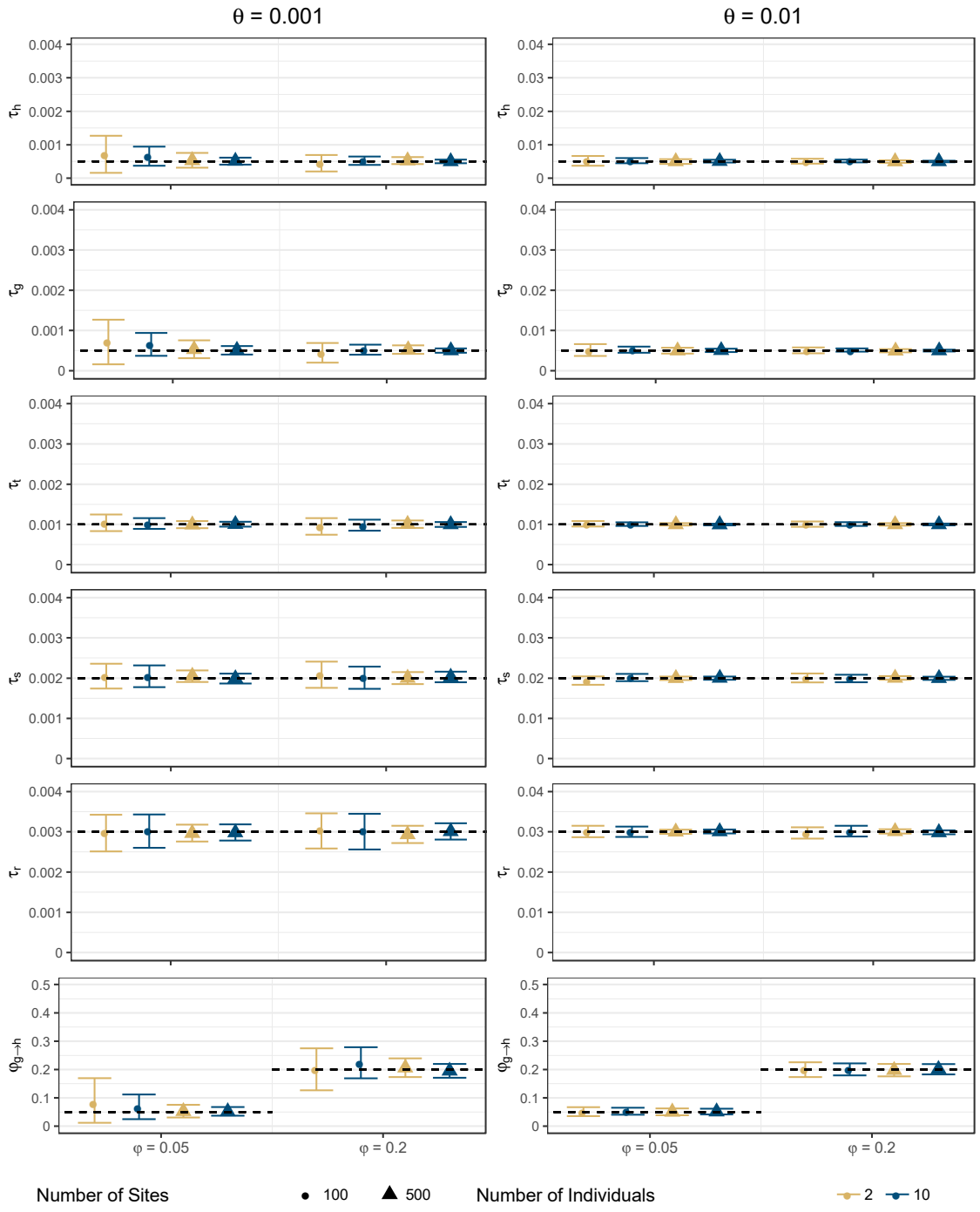

Figure S4 – Divergence Time and Introgression Probability Estimates under the MSci with Introgression between Non-Sister Lineages. Node labels and divergence times correspond to Fig. 1b.

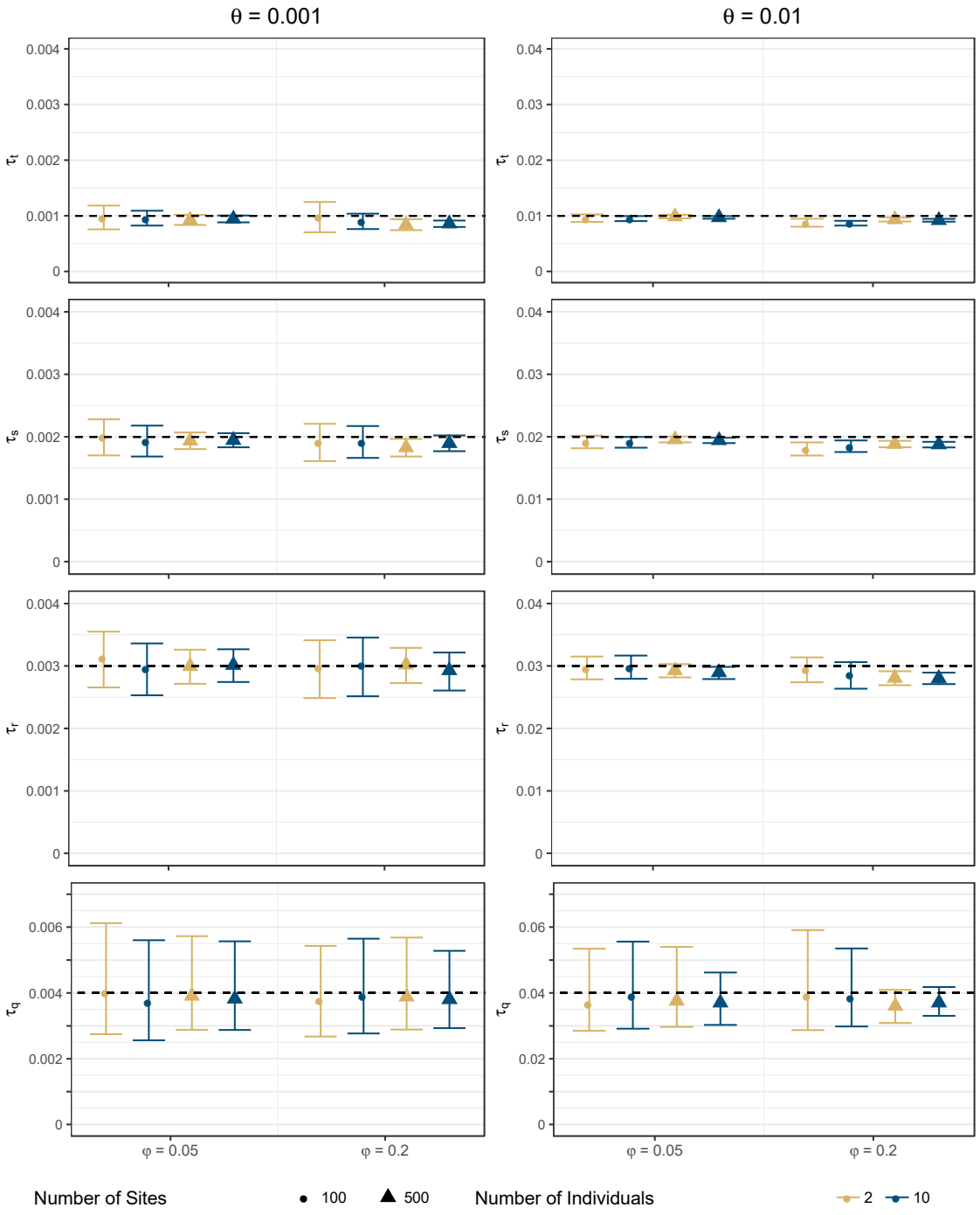

Figure S5 – Divergence Time Estimates under MSC with Introgression from a Ghost Lineage into an Ingroup Species. Node labels and divergence times correspond to Fig. 1c.

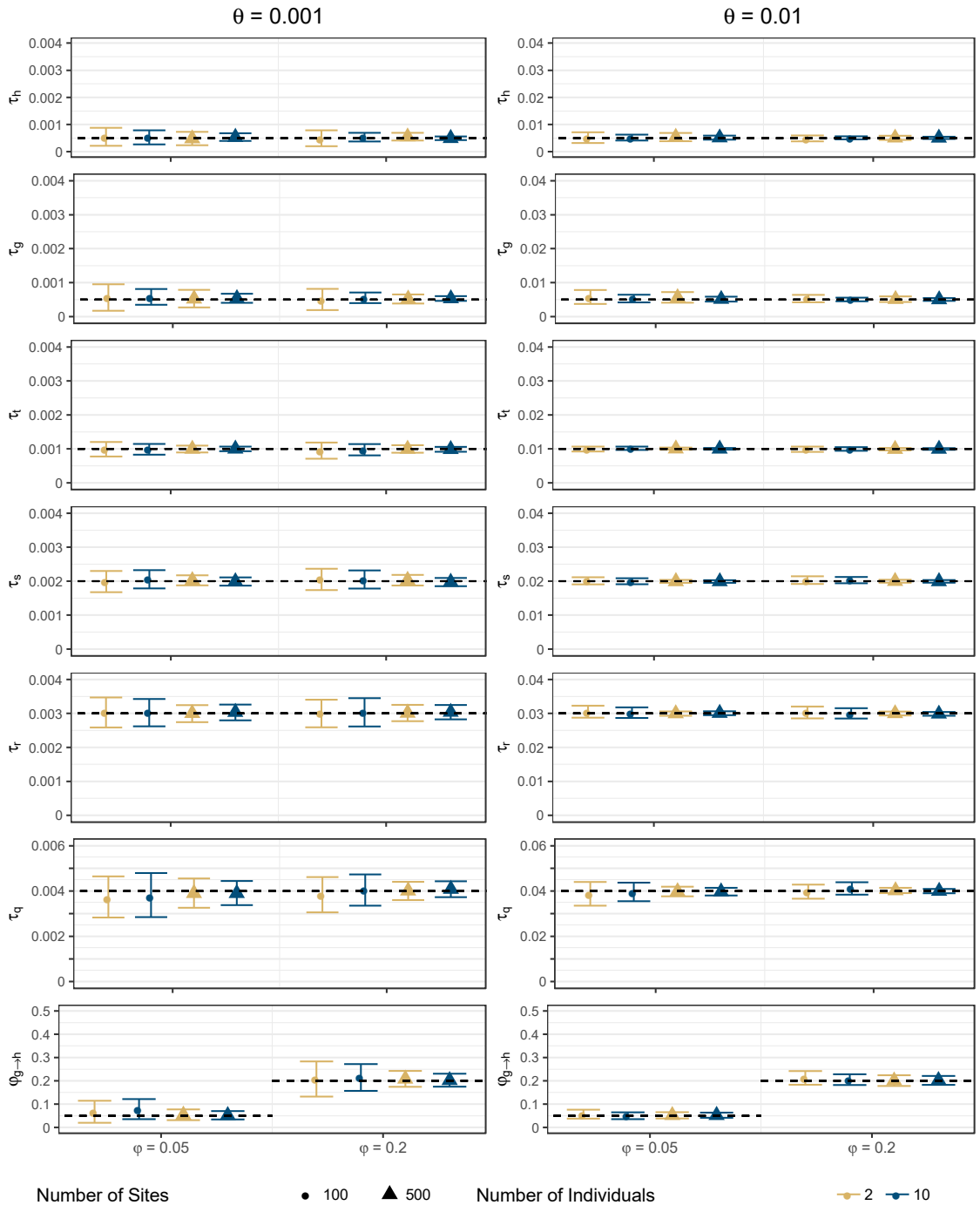

Figure S6 – Divergence Time and Introgression Probability Estimates under the MSci with Introgression from a Ghost Lineage into an Ingroup Species. Node labels and divergence times correspond to Fig. 1c.

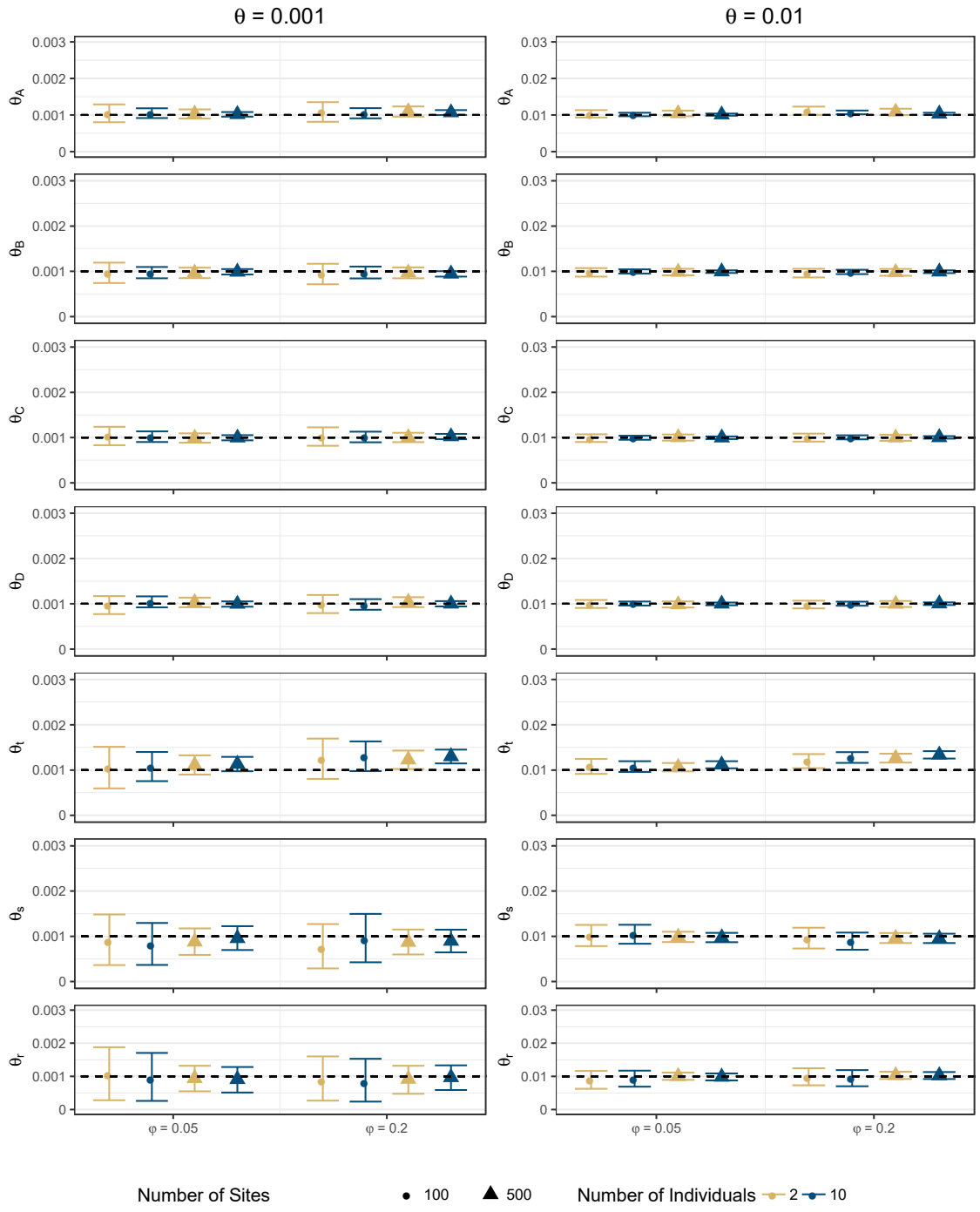

Figure S7 – Population Size Estimates under the MSC with Introgression between Sister Lineages. Node labels and populations sizes correspond to Fig. 1a.

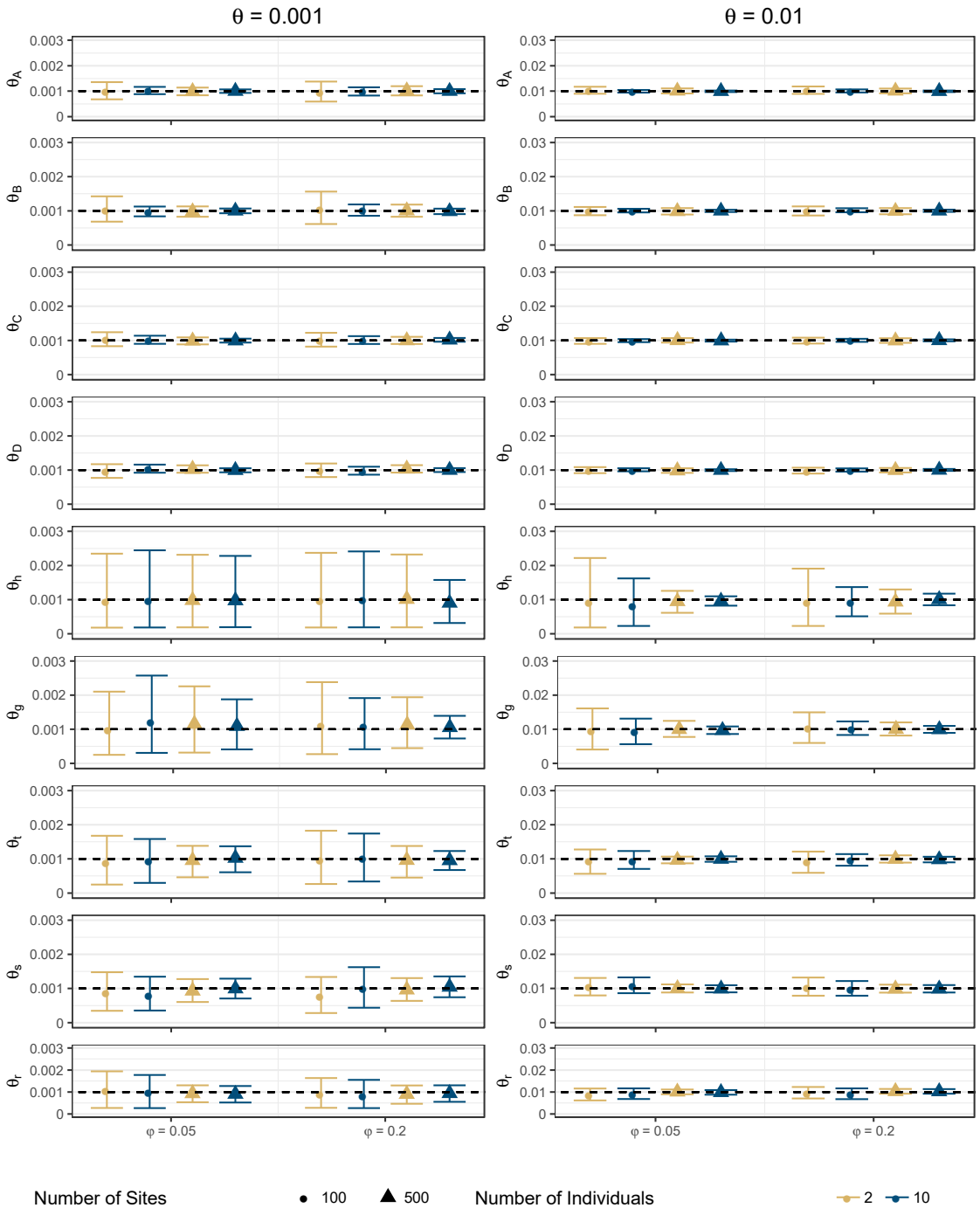

Figure S8 – Population Size Estimates under the MSci with Introgression between Sister Lineages. Node labels and population sizes correspond to Fig. 1a.

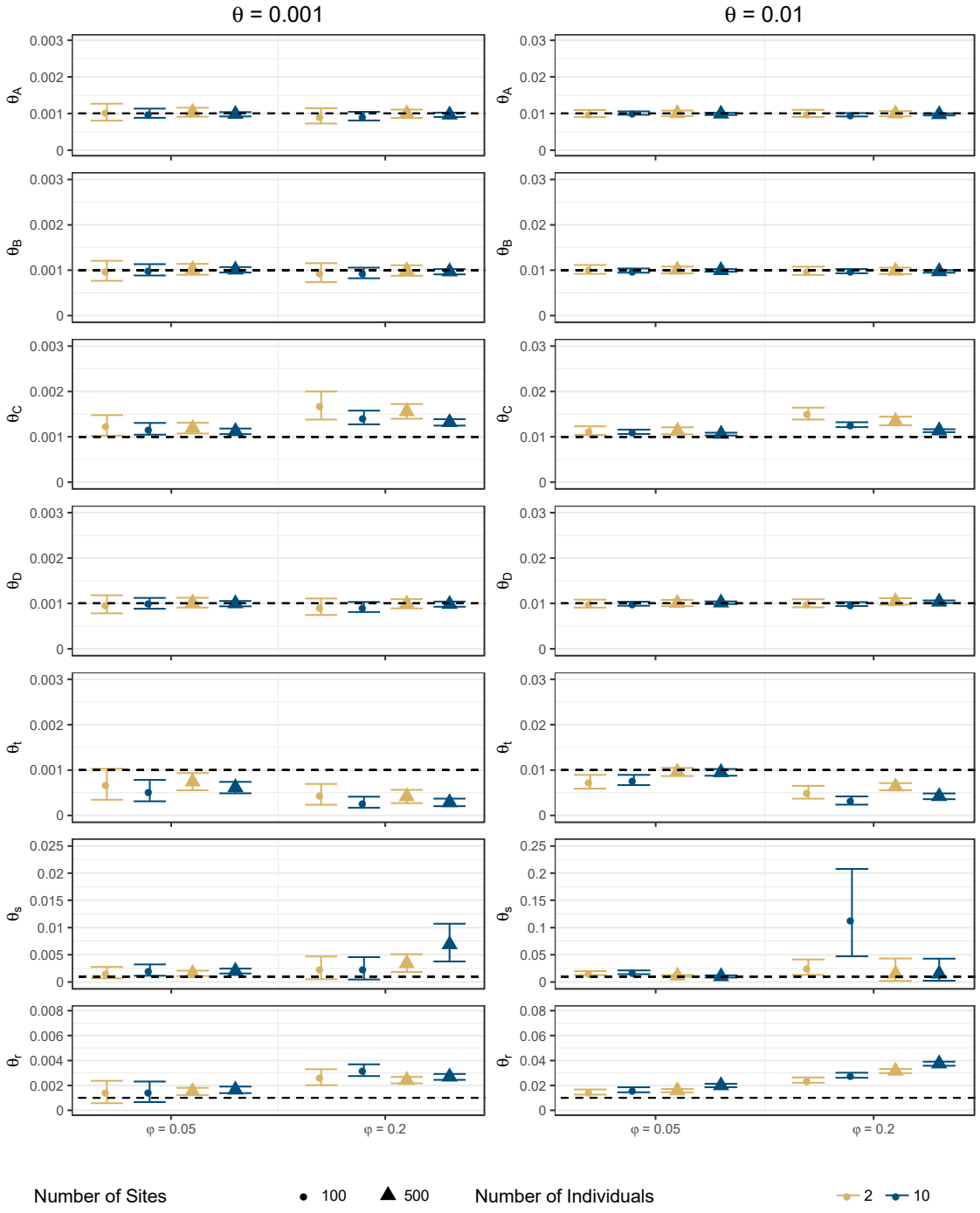

Figure S9 – Population Size Estimates under the MSC with Introgression between Non-Sister Lineages. Node labels and population sizes correspond to Fig. 1b.

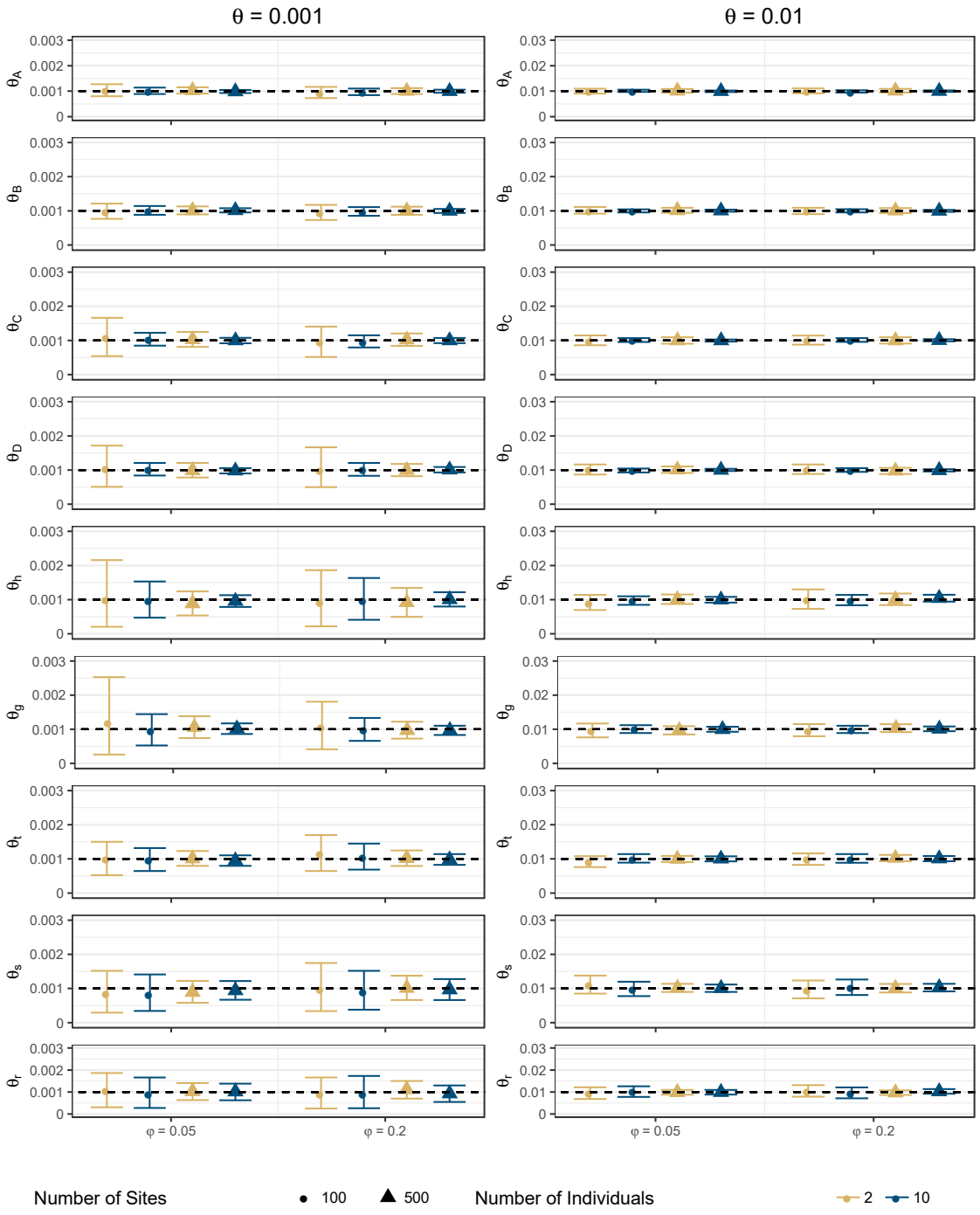

Figure S10 – Population Size Estimates under the MSci with Introgression between Non-Sister Lineages. P Node labels and population sizes correspond to Fig. 1b.

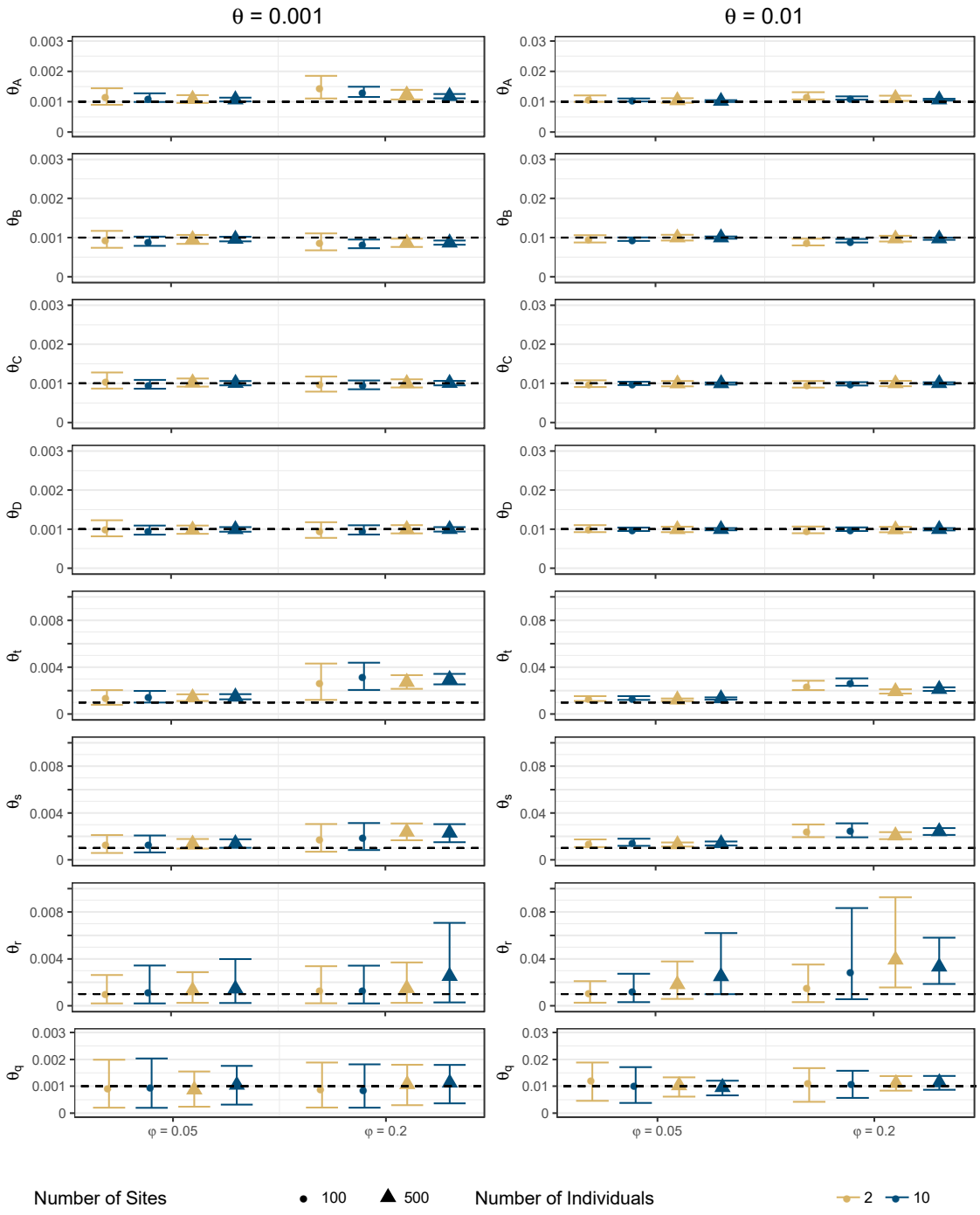

Figure S11 – Population Size Estimates under the MSC with Introgression from a Ghost Lineage to an Ingroup Species. Node labels and population sizes correspond to Fig. 1c.

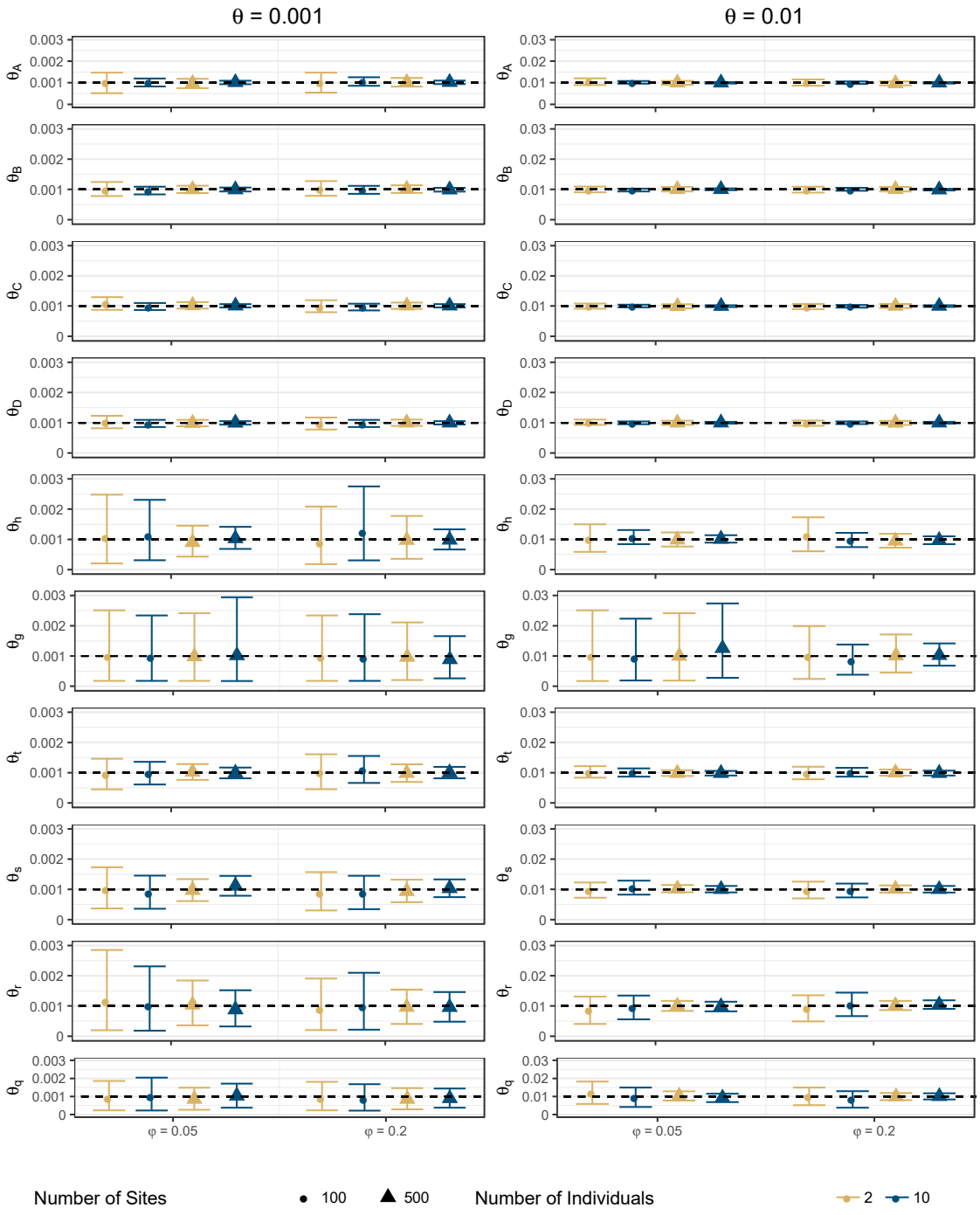

Figure S12 – Population Size Estimates under the MSci with Introgression from a Ghost Lineage to an Ingroup Species. Node labels and population sizes correspond to Fig. 1c.

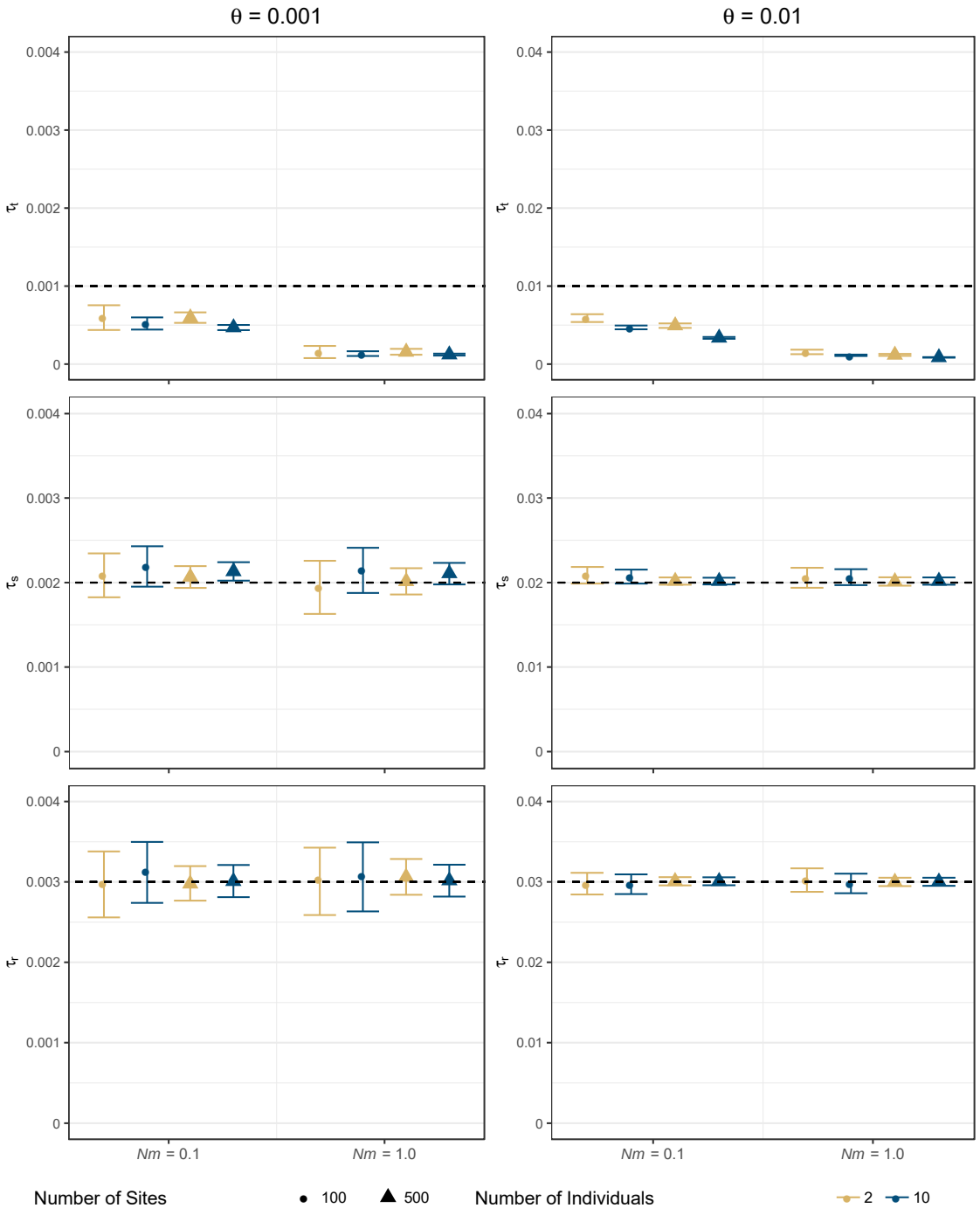

Figure S13 – Divergence Time Estimates under the MSC when there is migration between Sister Lineages. Node labels and divergence times correspond to Fig. 1a. *Nm* is the population immigration rate.

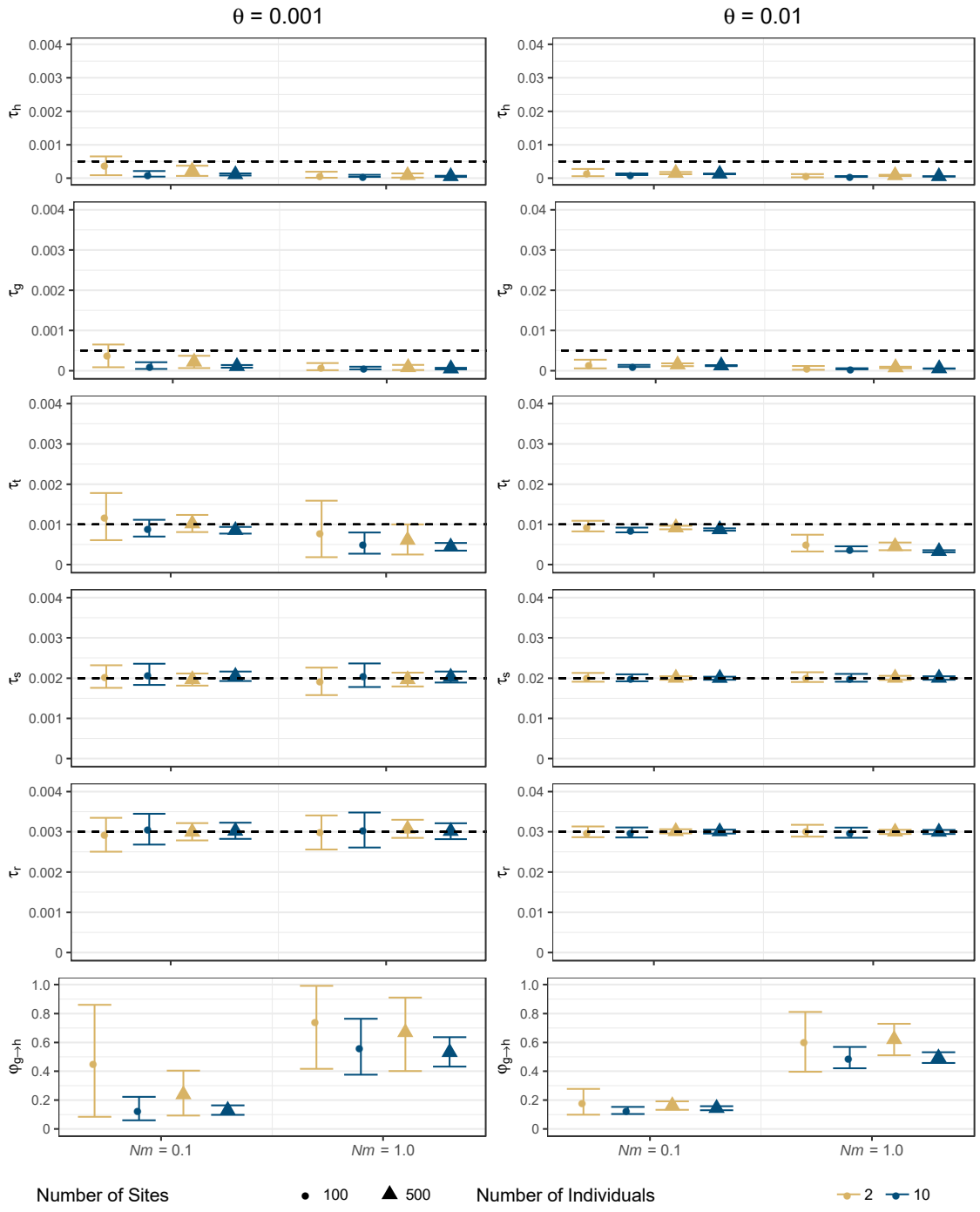

Figure S14 – Divergence Time and Introgression Probability Estimates under the MSci when there is migration between Sister Lineages. Node labels and divergence times correspond to Fig. 1a. *Nm* is the population immigration rate, so there is no true value for φ_h.

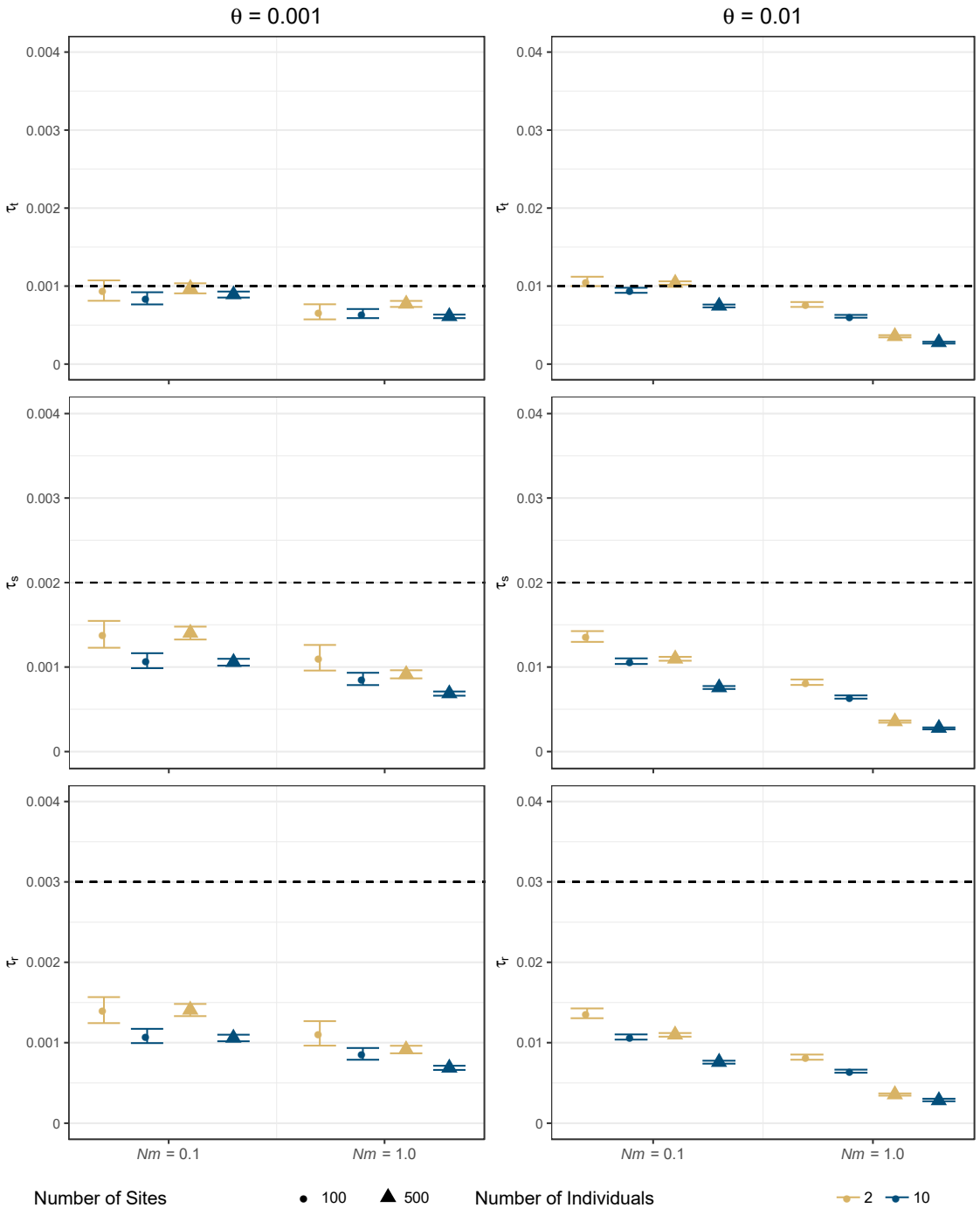

Figure S15 – Divergence Time Estimates under the MSC when there is migration between Non-Sister Lineages. Node labels and divergence times correspond to Fig. 1b.

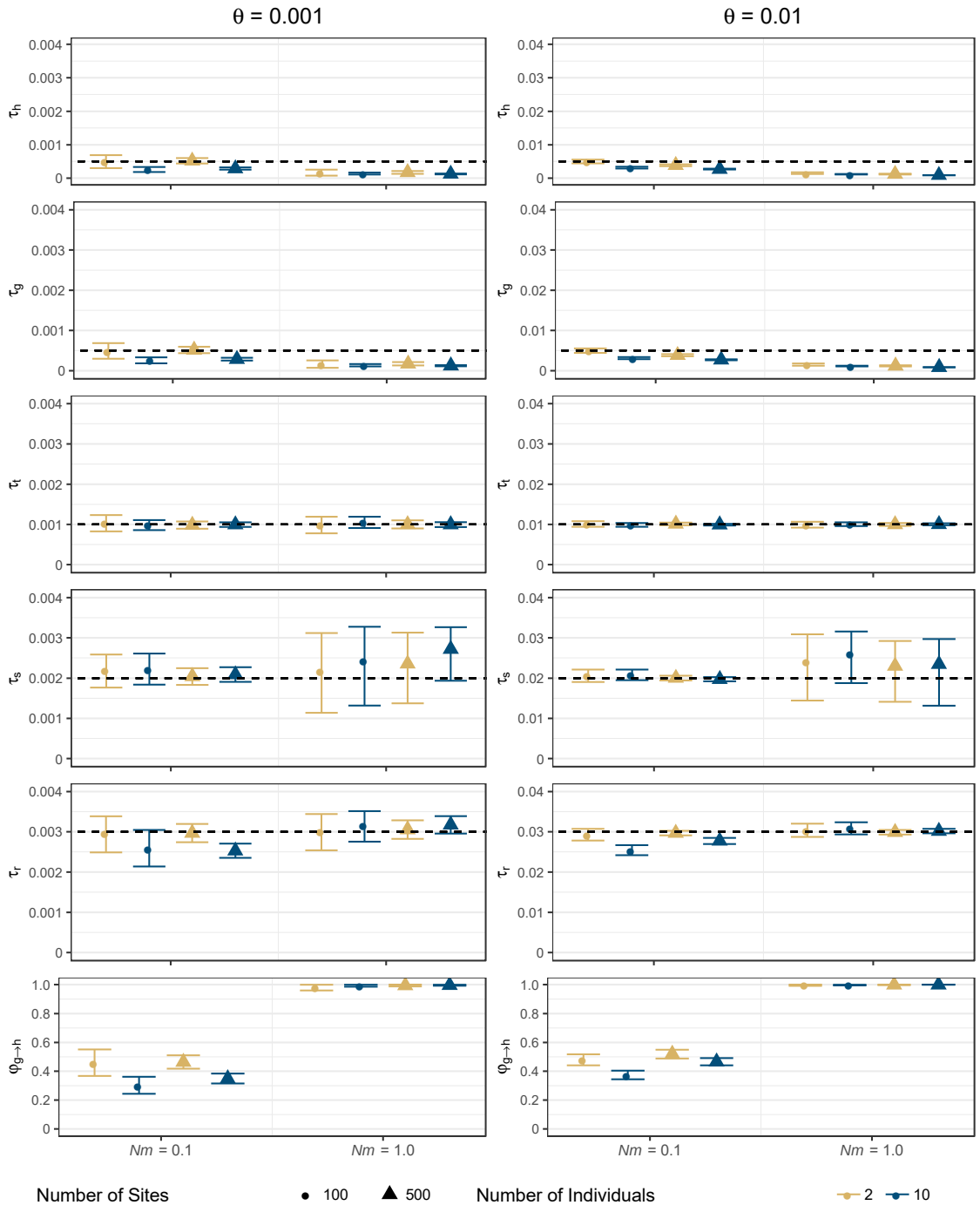

Figure S16 – Divergence Time and Introgression Probability Estimates under the MSci when there is migration between Non-Sister Lineages. Node labels and divergence times correspond to Fig. 1b.

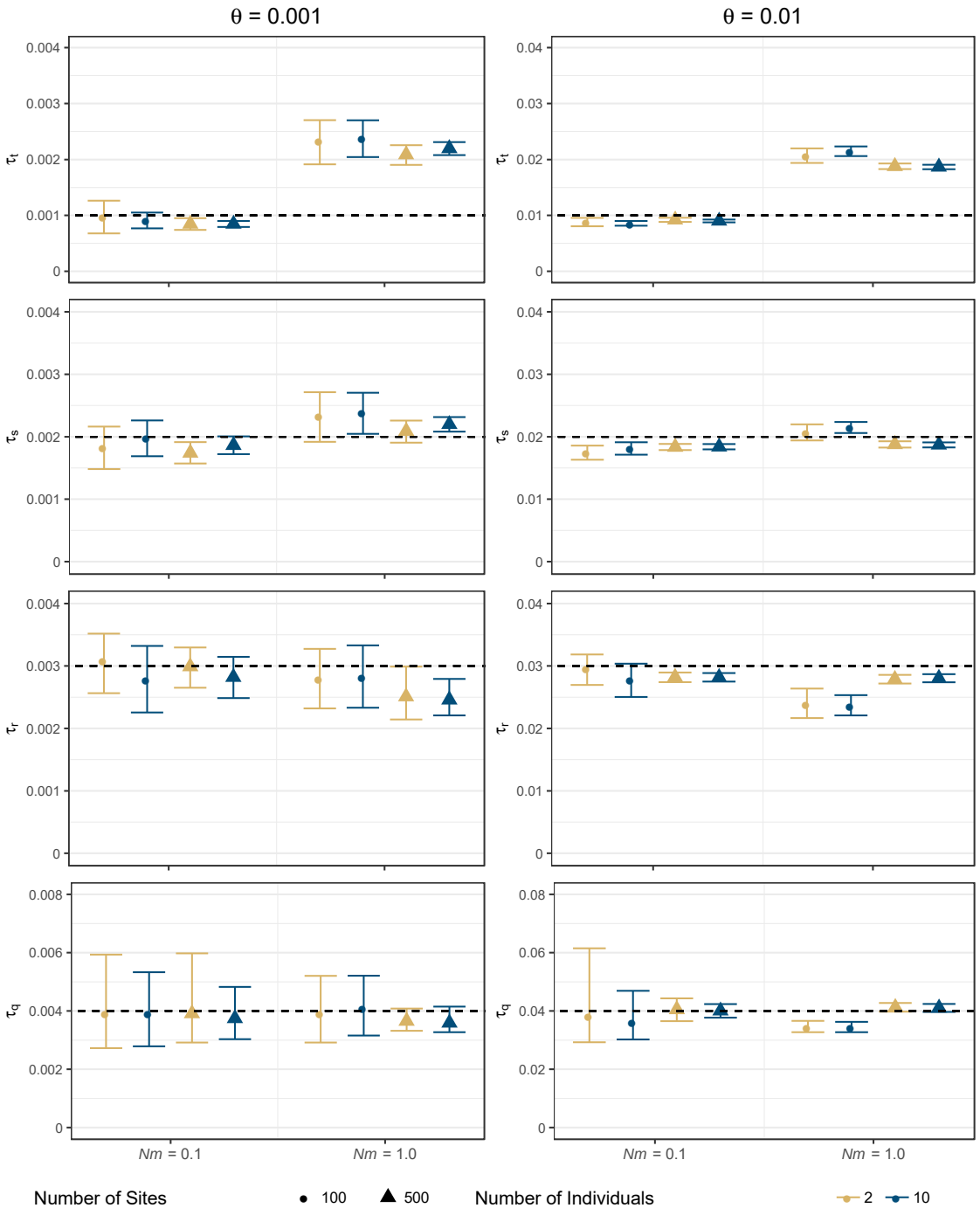

Figure S17 – Divergence Time Estimates under the MSC When there is Migration from a Ghost Lineage to an Ingroup Species. Node labels and divergence times correspond to Fig. 1c.

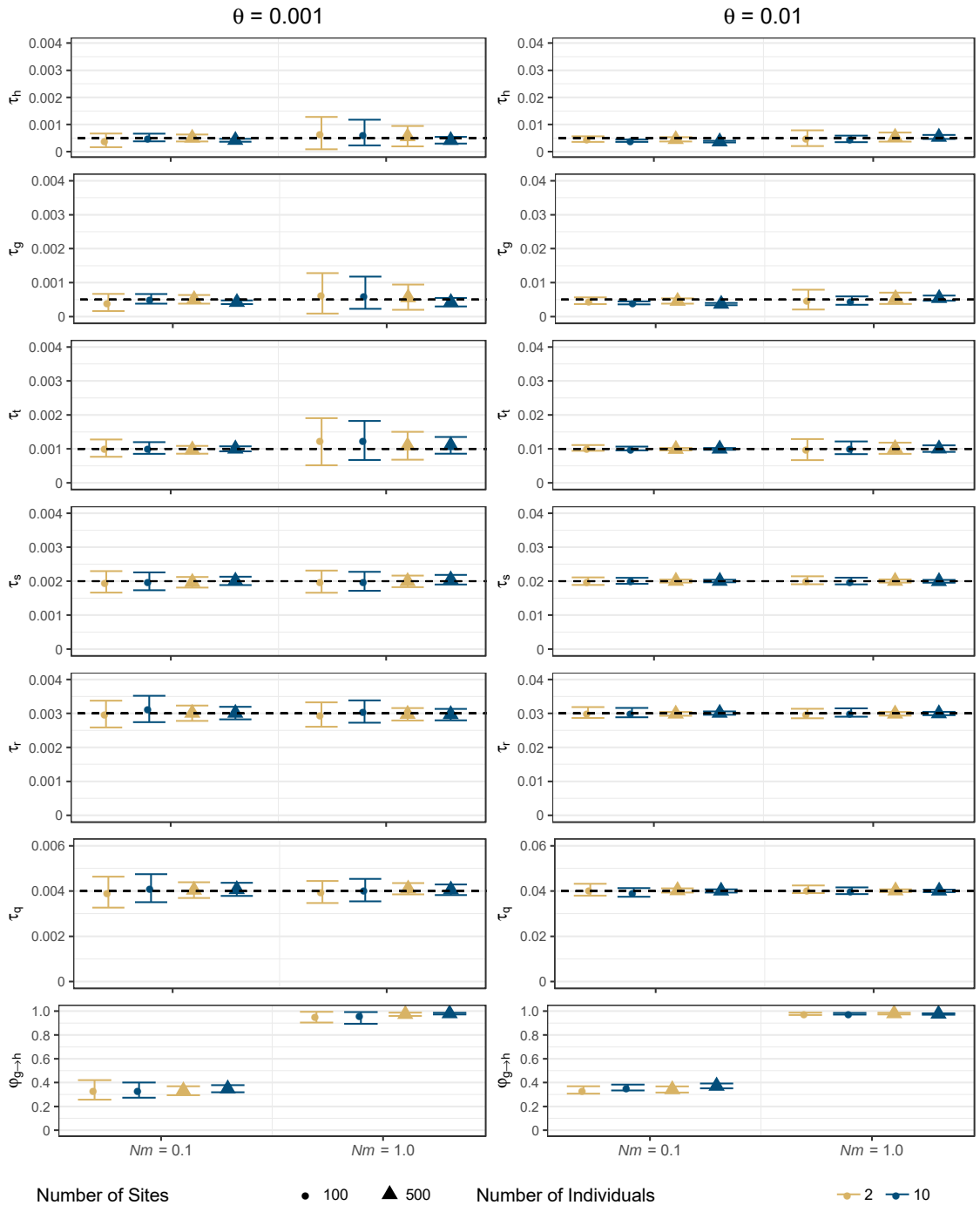

Figure S18 – Divergence Time and Introgression Probability Estimates under the MSci When there is Migration from a Ghost Lineage to an Ingroup Species. Node labels and divergence times correspond to Fig. 1c.

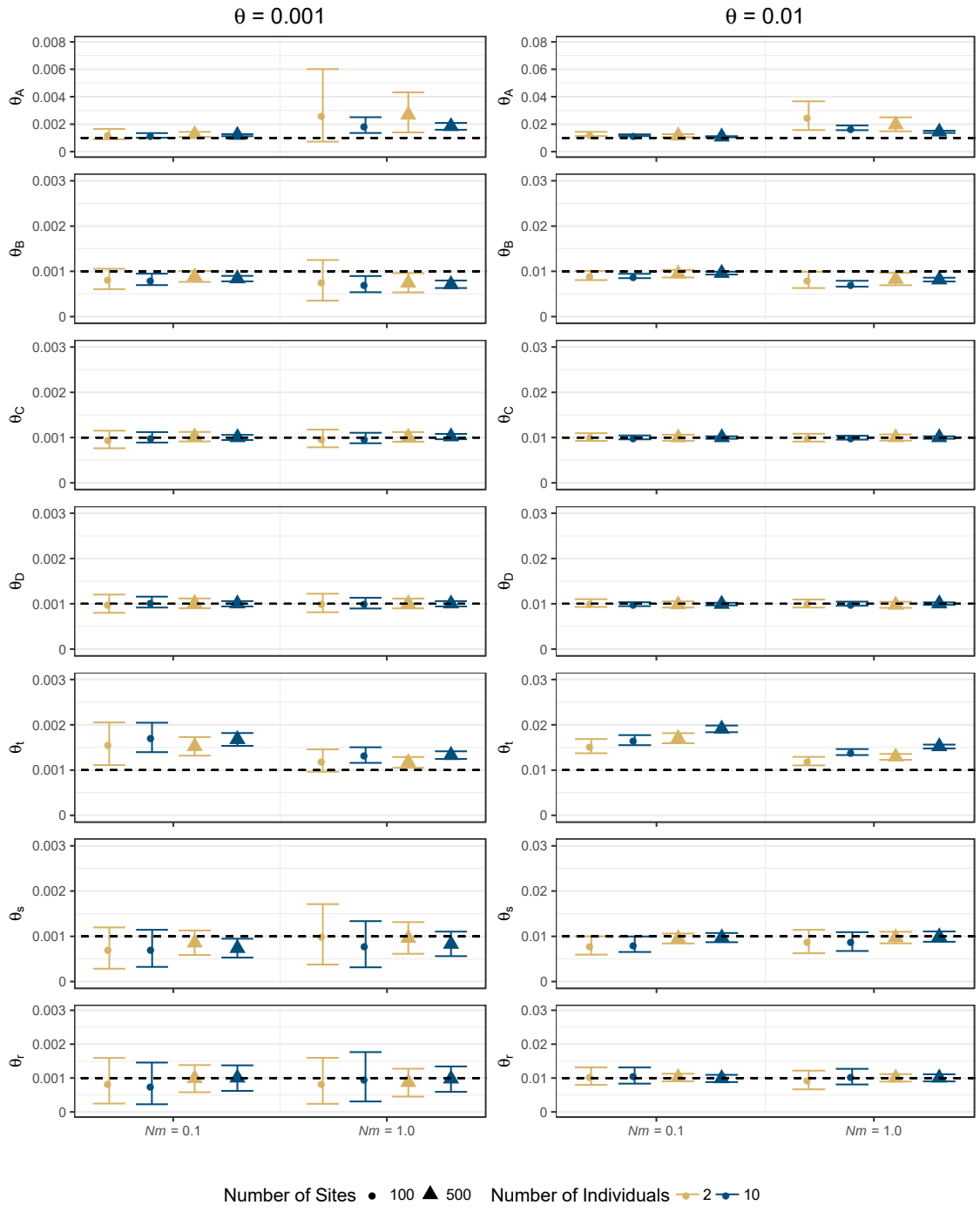

Figure S19 – Population Size Estimates under the MSC When There is Migration Between Sister Lineages. Node labels and populations sizes correspond to Fig. 1a.

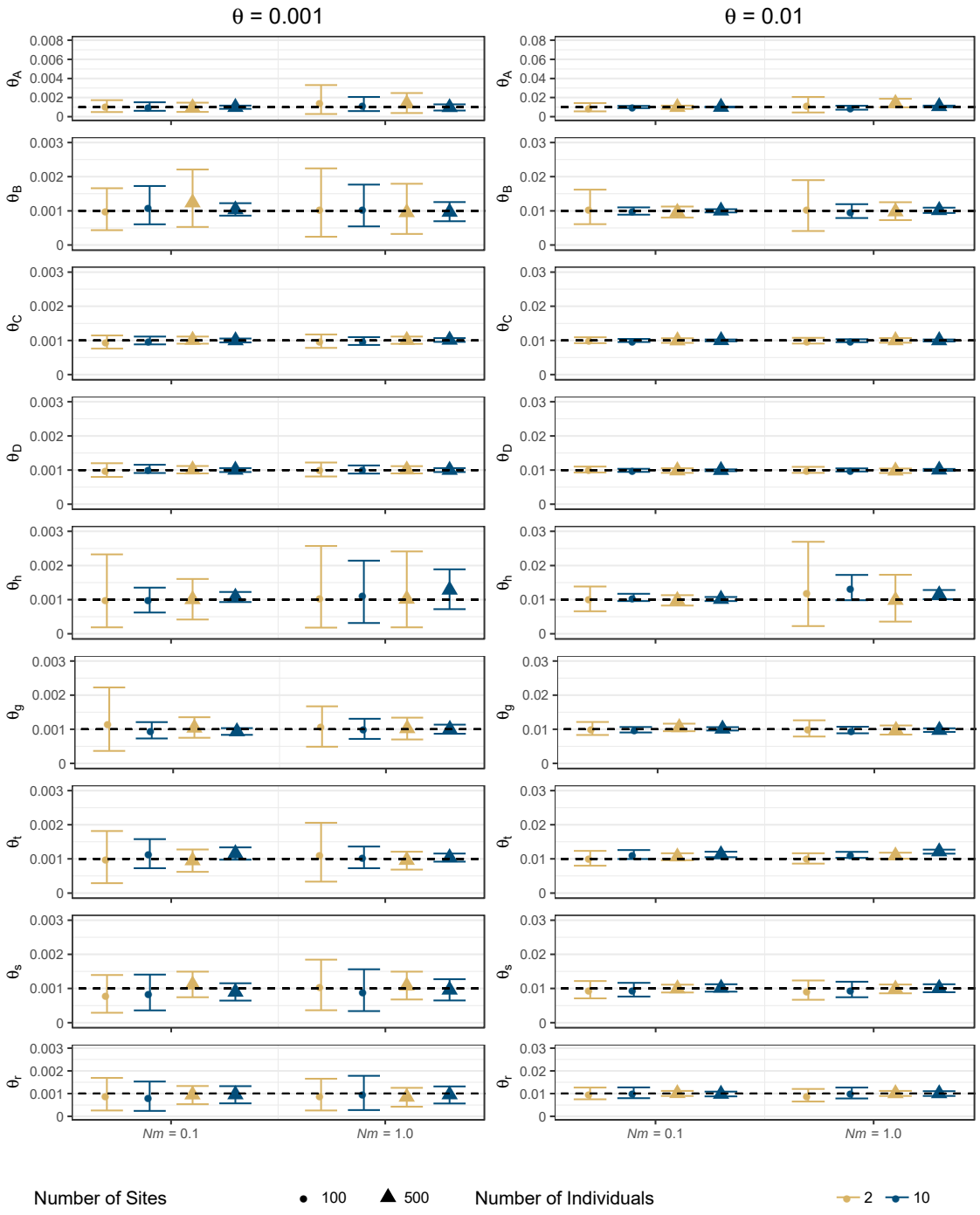

Figure S20 – Population Size Estimates under the MSci When There is Migration Between Sister Lineages. Node labels and population sizes correspond to Fig. 1a.

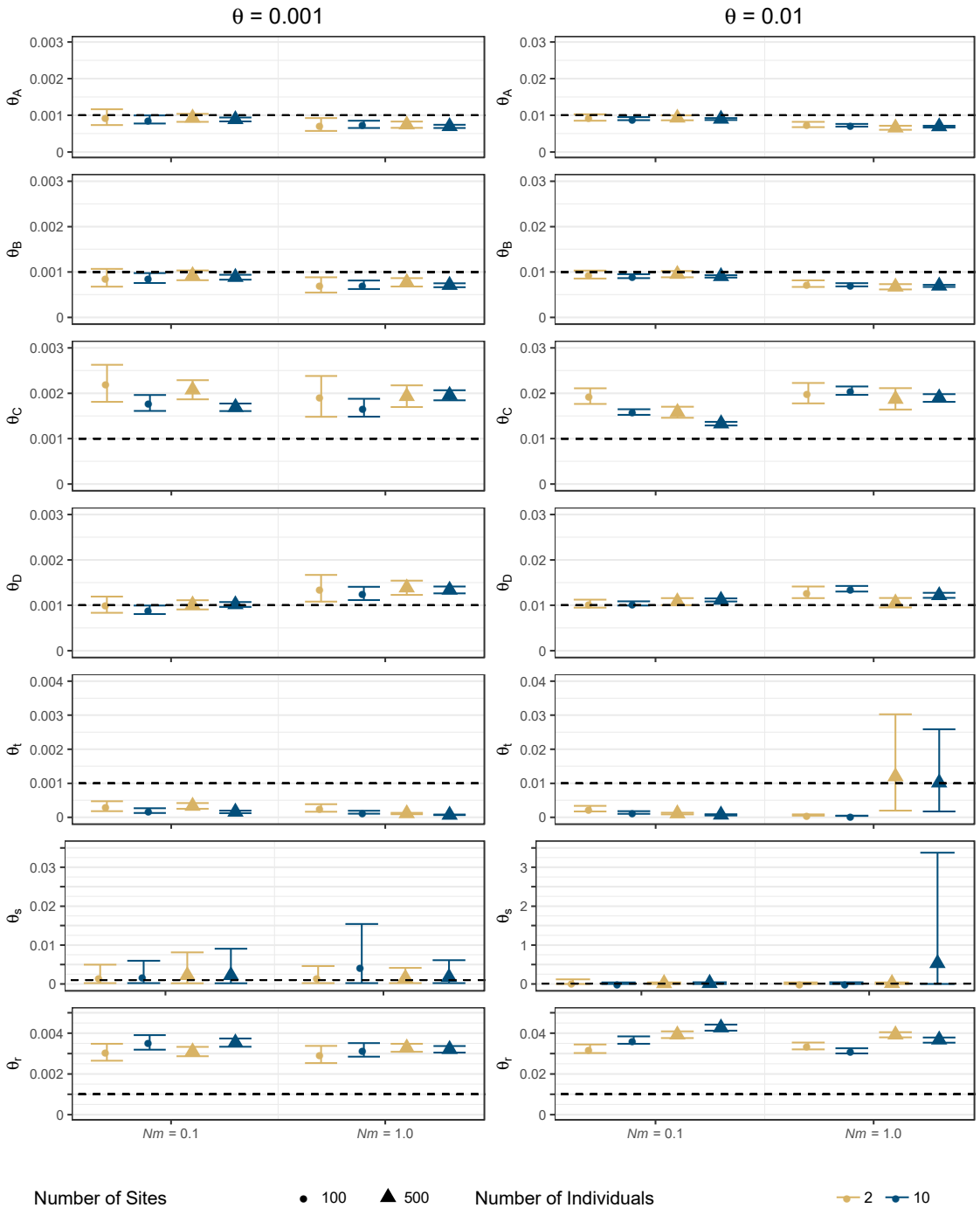

Figure S21 – Population Size Estimates under the MSC When There is Migration Between Non-Sister Lineages. Node labels and population sizes correspond to Fig. 1b.

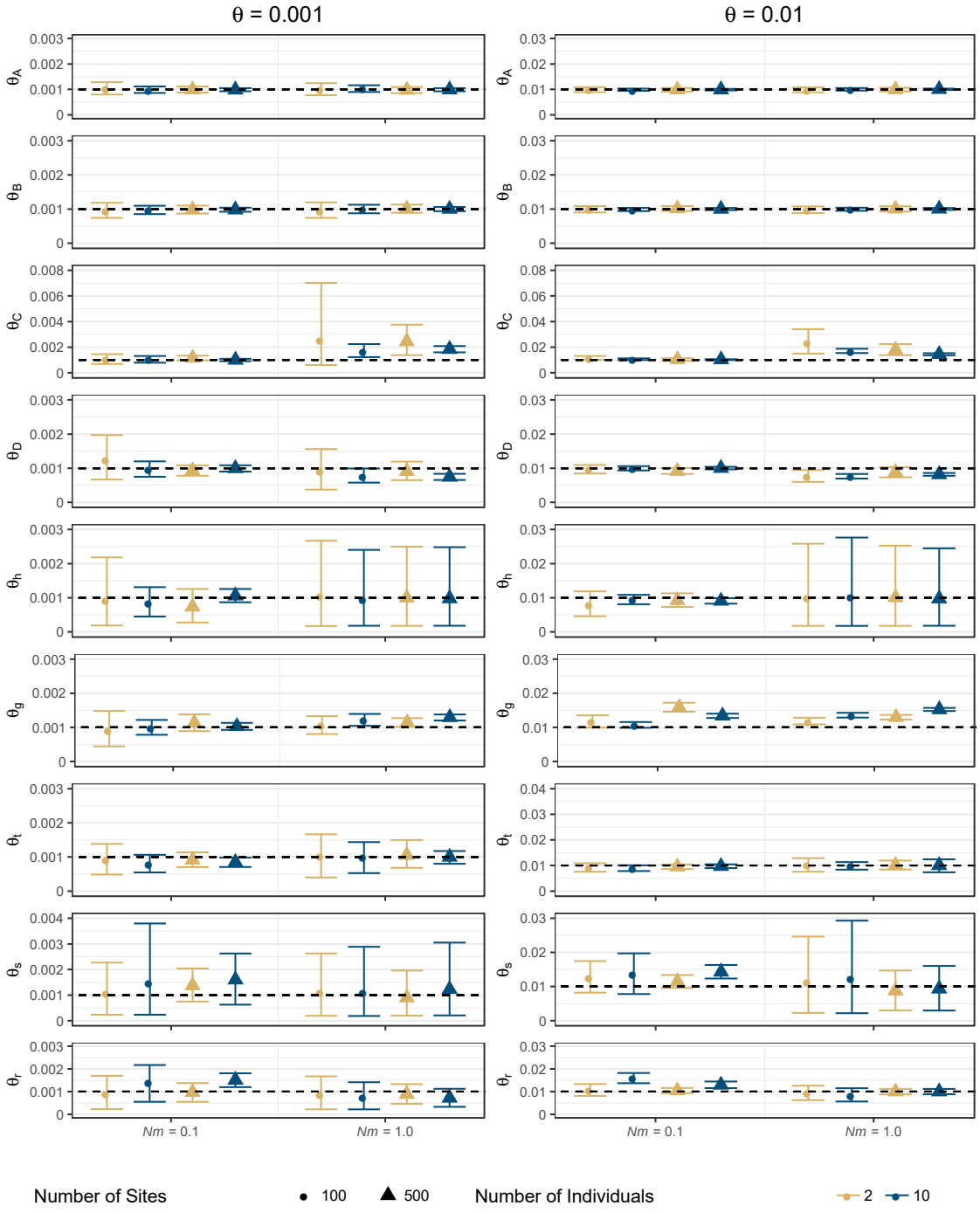

Figure S22 – Population Size Estimates under the MSci When There is Migration Between Non-Sister Lineages. Node labels and population sizes correspond to Fig. 1b.

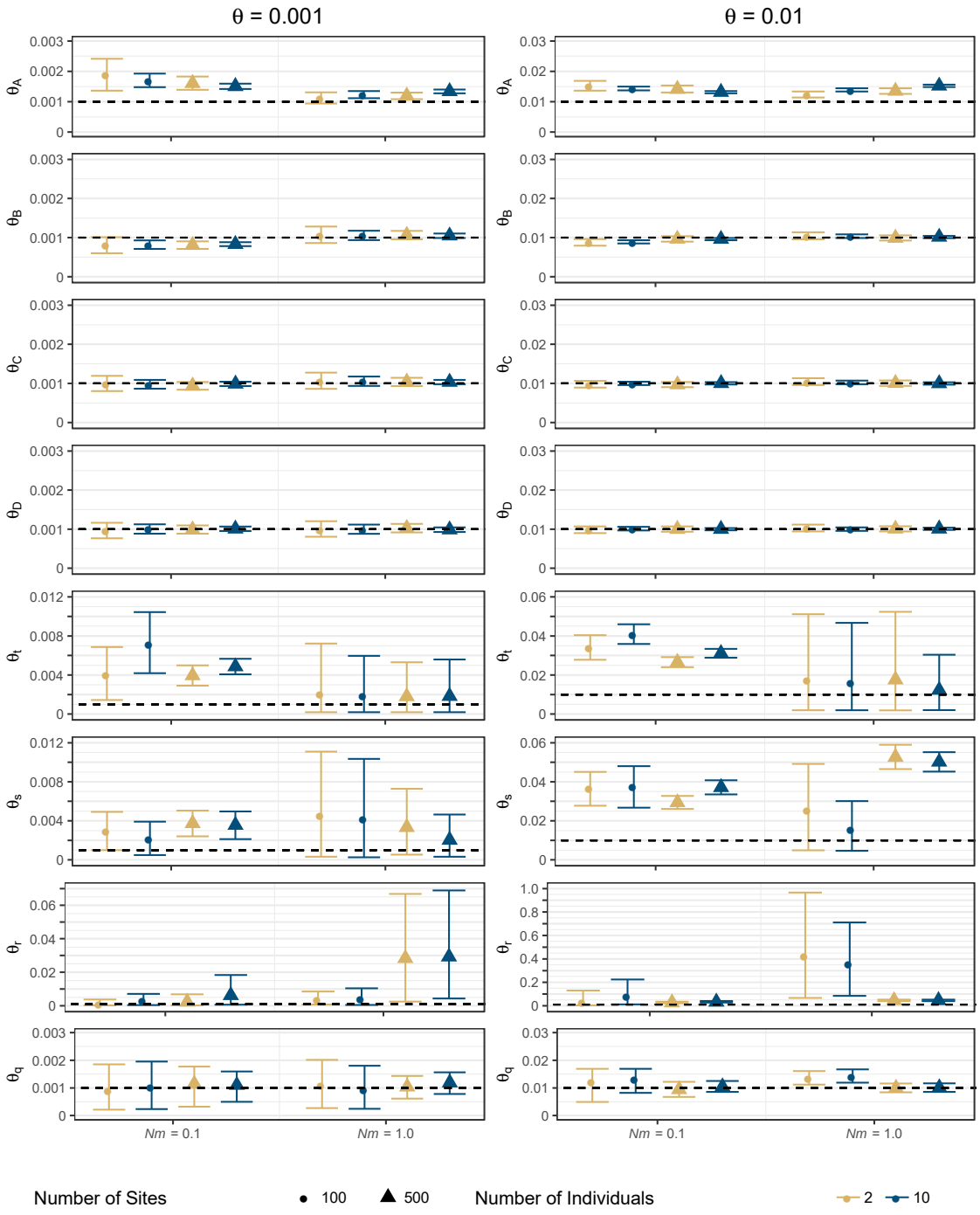

Figure S23 – Population Size Estimates under the MSC When There is Migration from a Ghost Lineages to an Ingroup Species. Node labels and population sizes correspond to Fig. 1c.

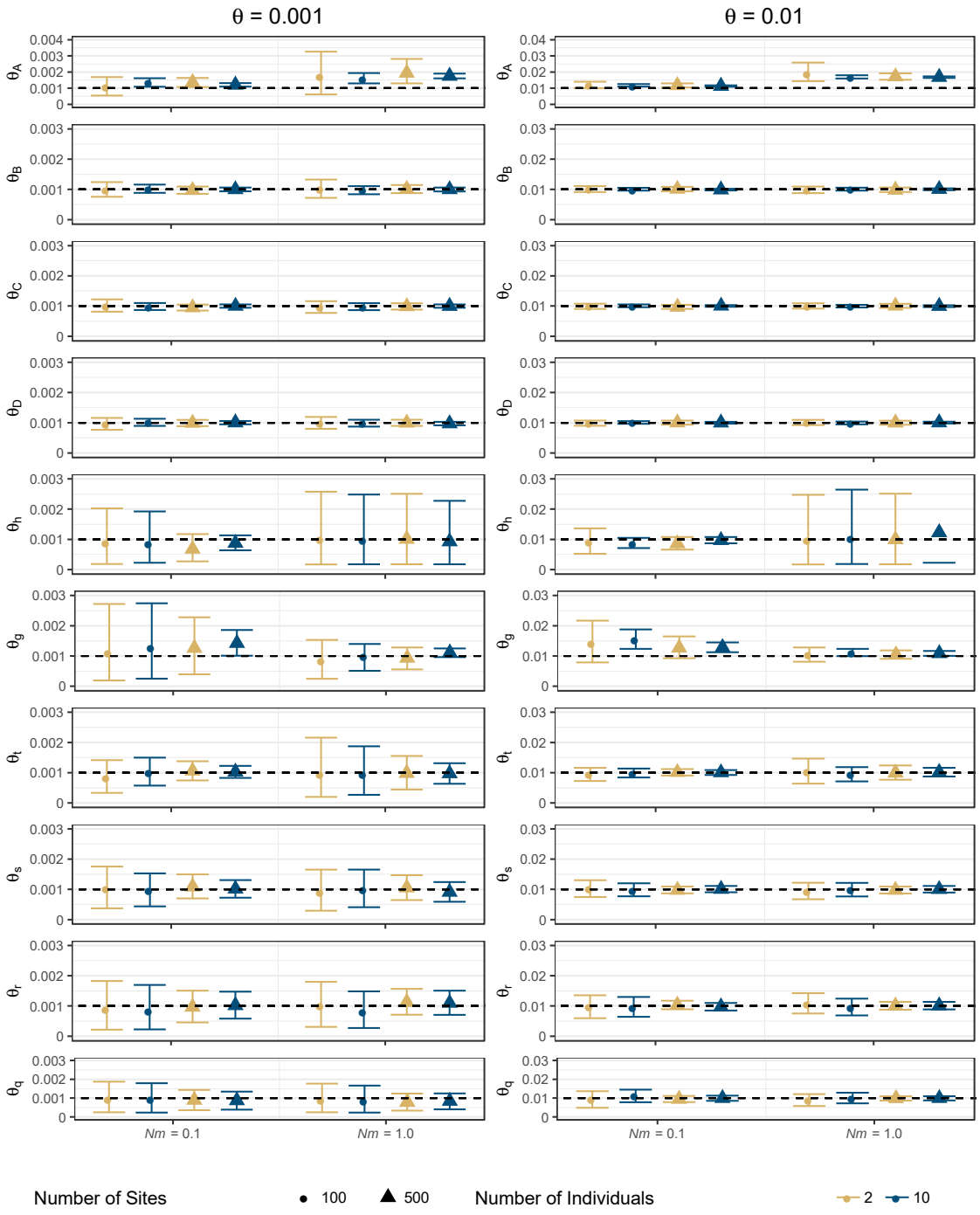

Figure S24 – Population Size Estimates under the MSci When There is Migration from a Ghost Lineages to an Ingroup Species. Node labels and population sizes correspond to Fig. 1c.

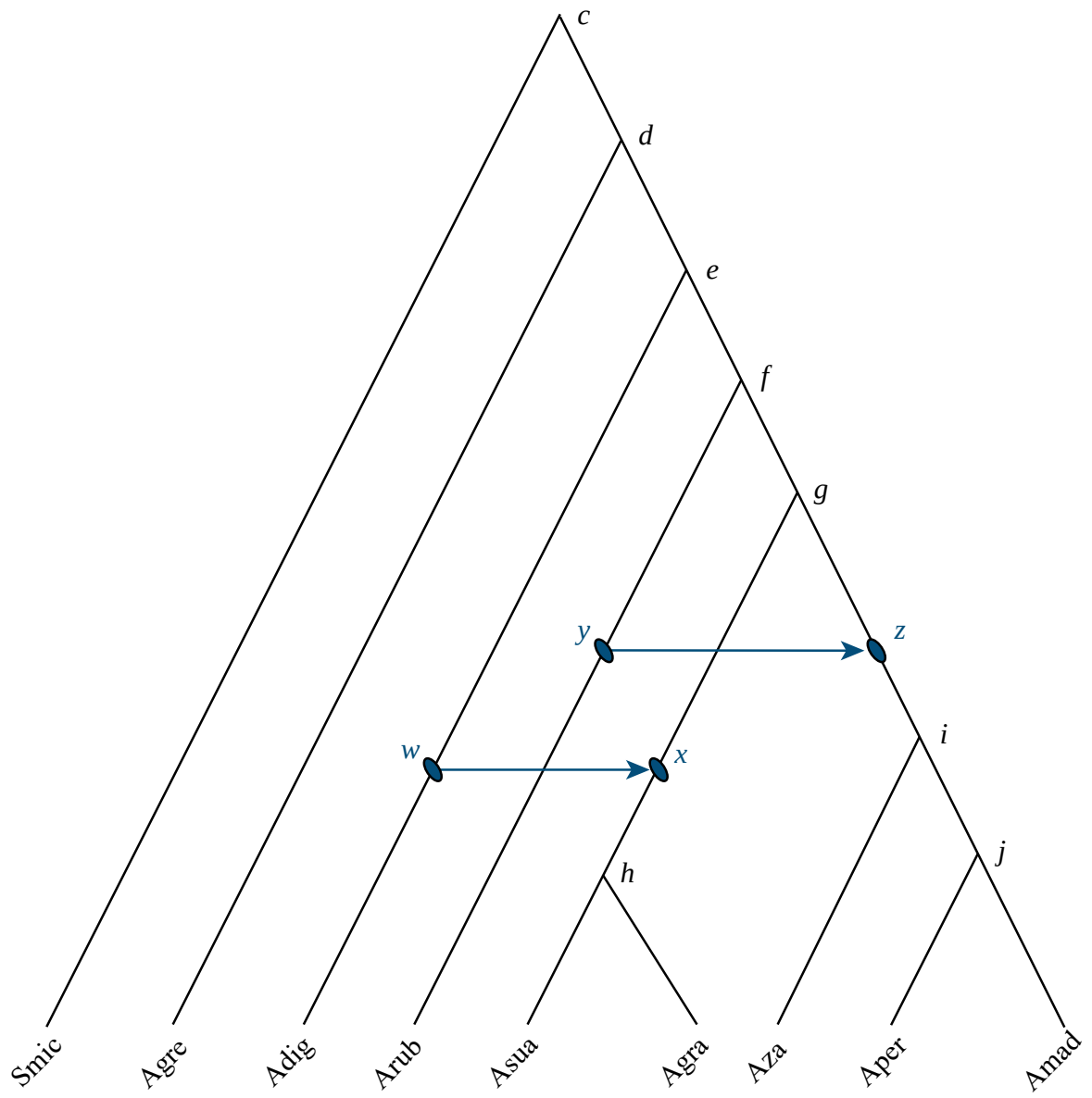

Figure S25 – Species Tree with Node Labels for *Adansonia* Assumed in BPP Analyses. Divergence time estimation used the inferred introgression events from PhyloNetworks. Node labels are used to identify parameters. Blue edges represent introgression events assumed in the MSci model.

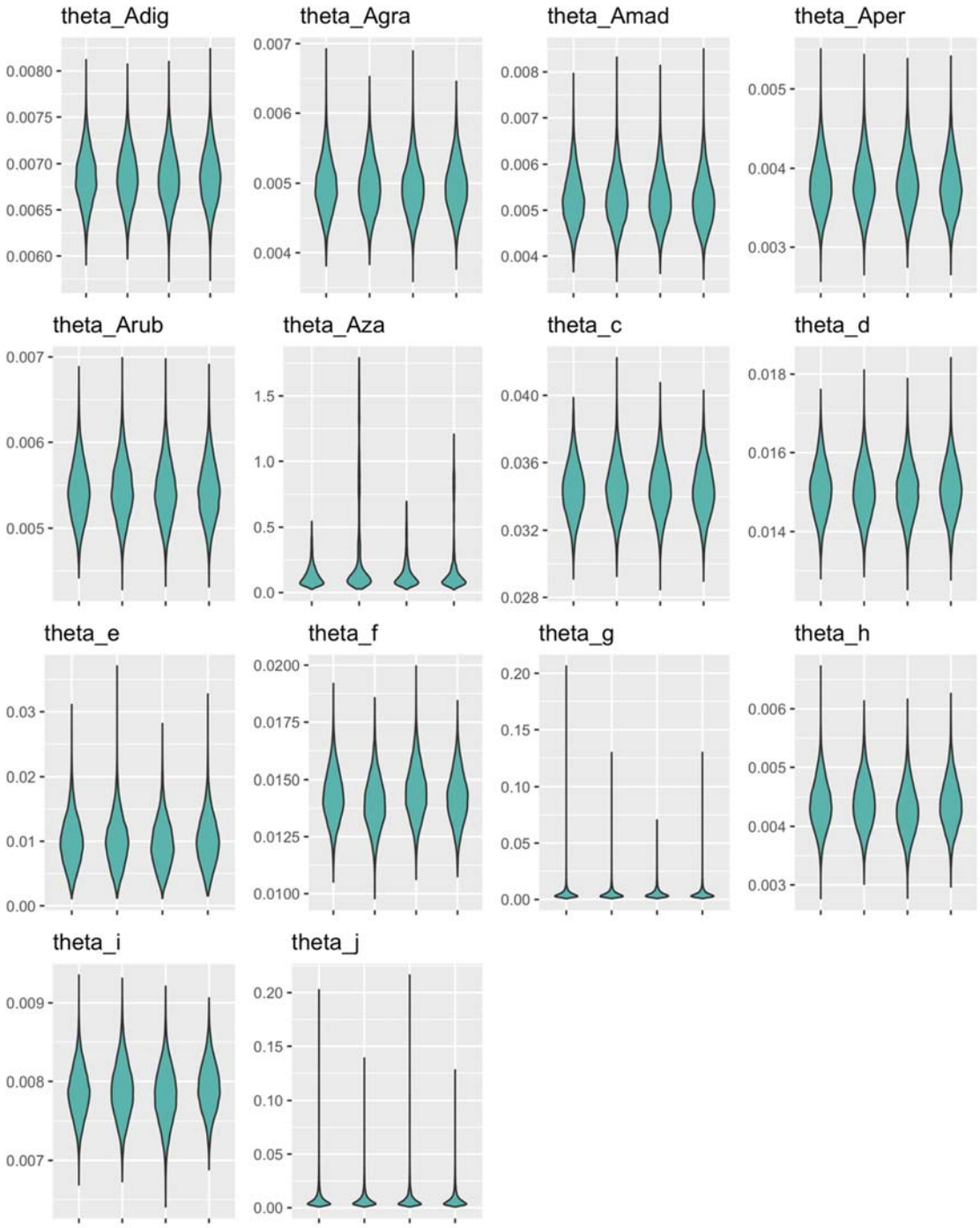

Figure S26 – Population Size Estimates for *Adansonia* under the MSC. The four violin plots represent the four individual chains used for analysis.

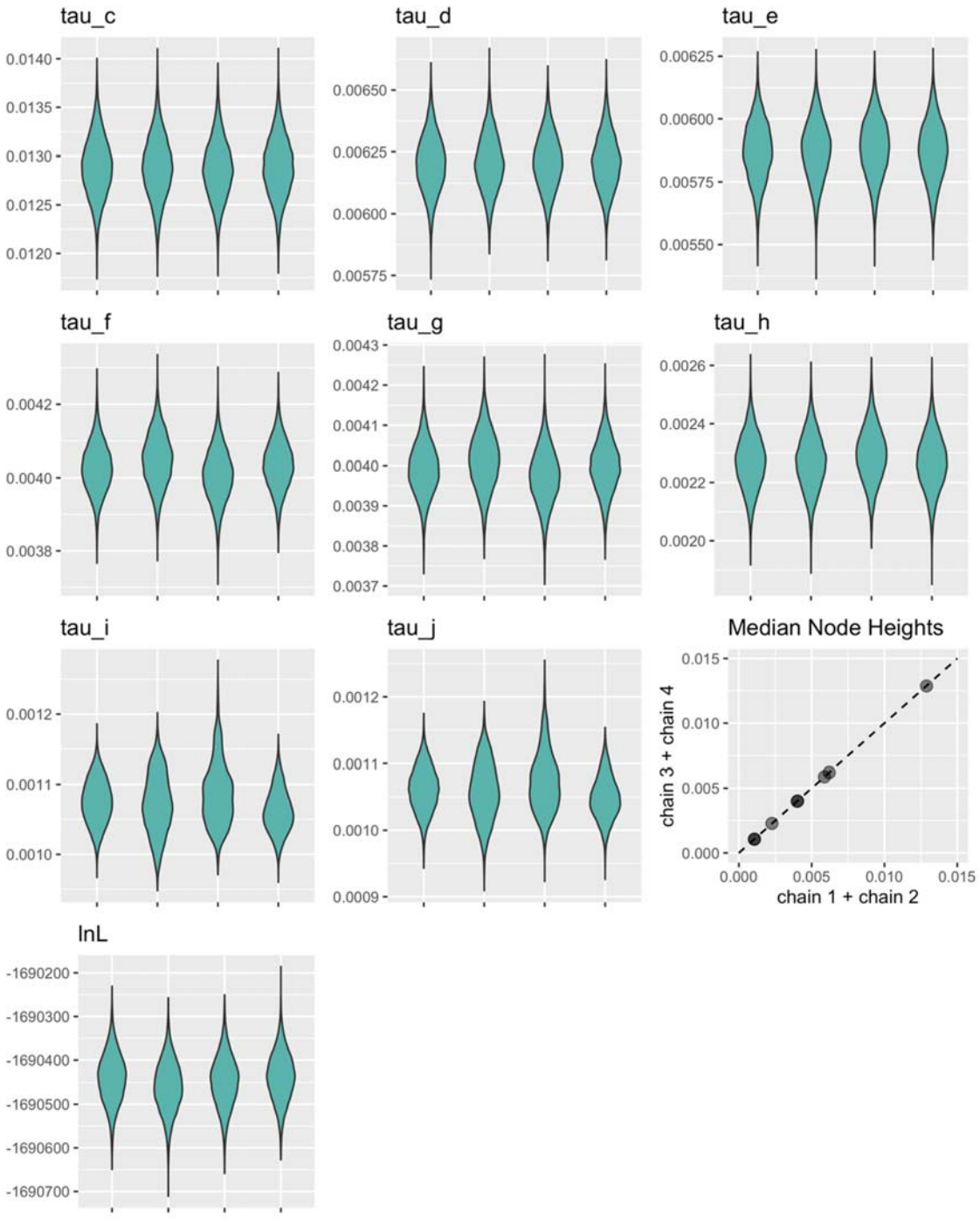

Figure S27 – Divergence Time Estimates for *Adansonia* under the MSC. The four violin plots represent the four individual chains used for analysis. A scatterplot of median node heights for τ implies convergence of the MCMC analyses.

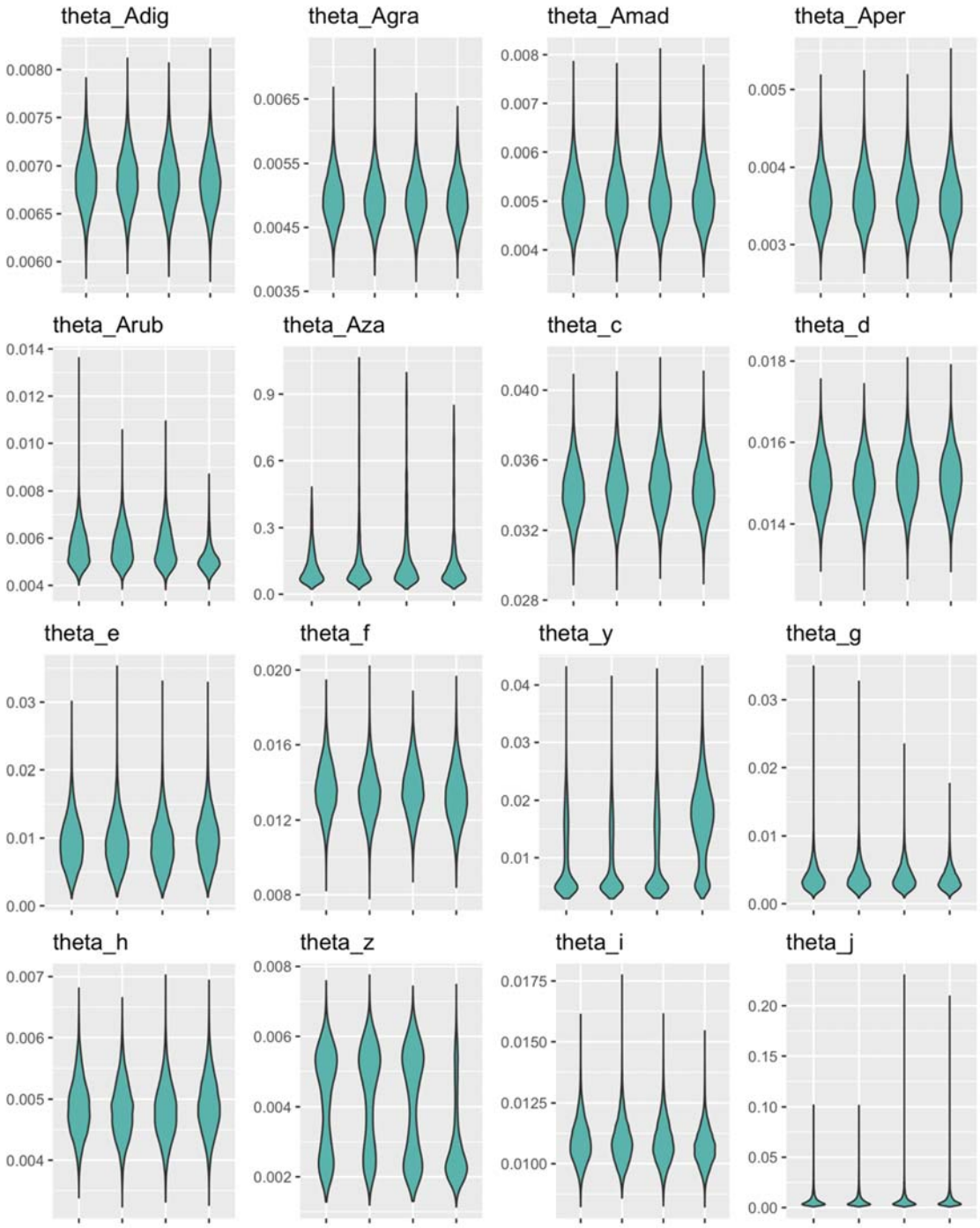

Figure S28 – Population Size Estimates for *Adansonia* under the MSci. The four violin plots represent the four individual chains used for analysis. Additional nodes for introgression events are represented here.

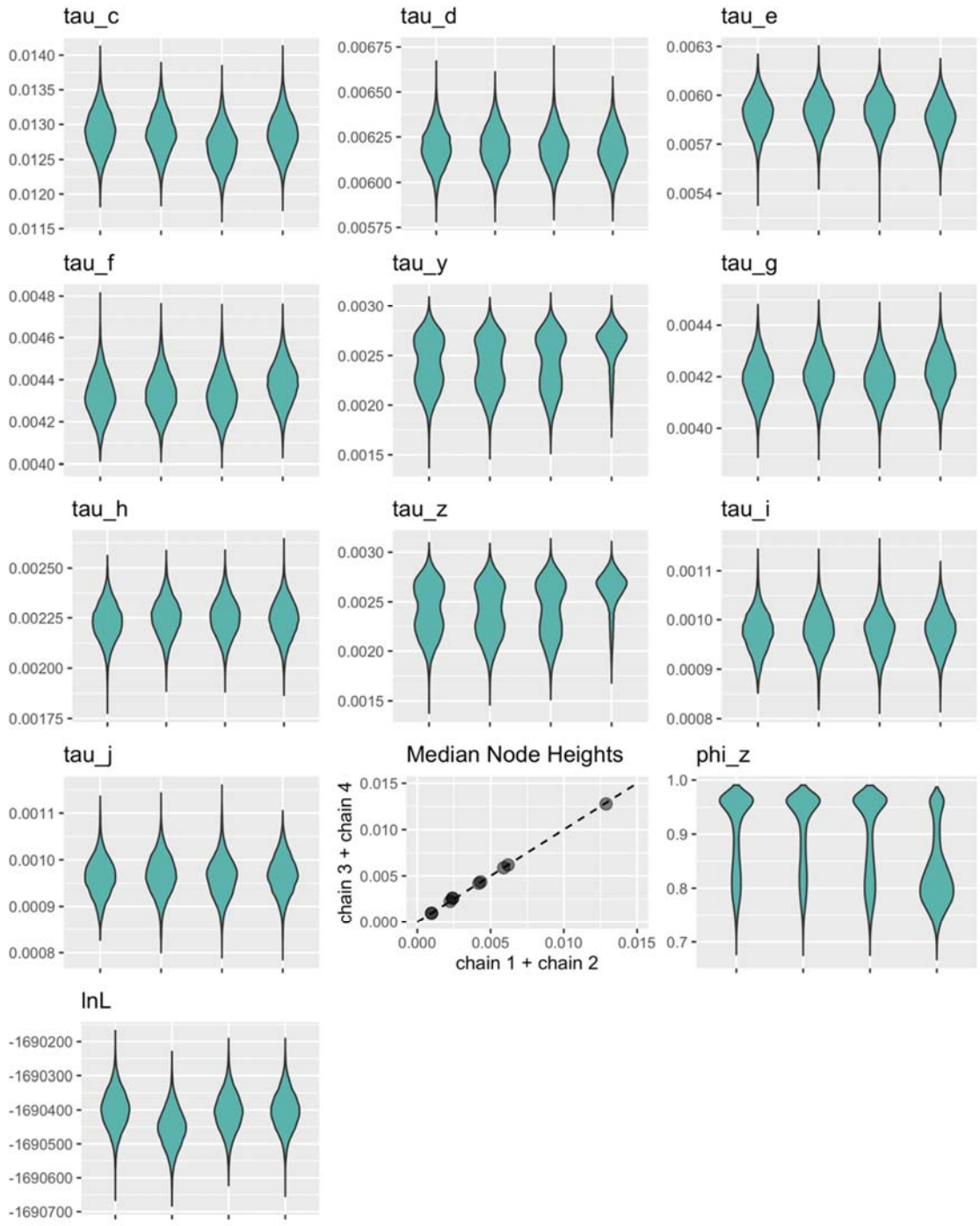

Figure S29 – Posterior Estimates of Divergence Times and Introgression Probability under the MSci for *Adansonia*. The four violin plots represent four replicate runs. The scatterplot of median node ages suggests consistency between runs.

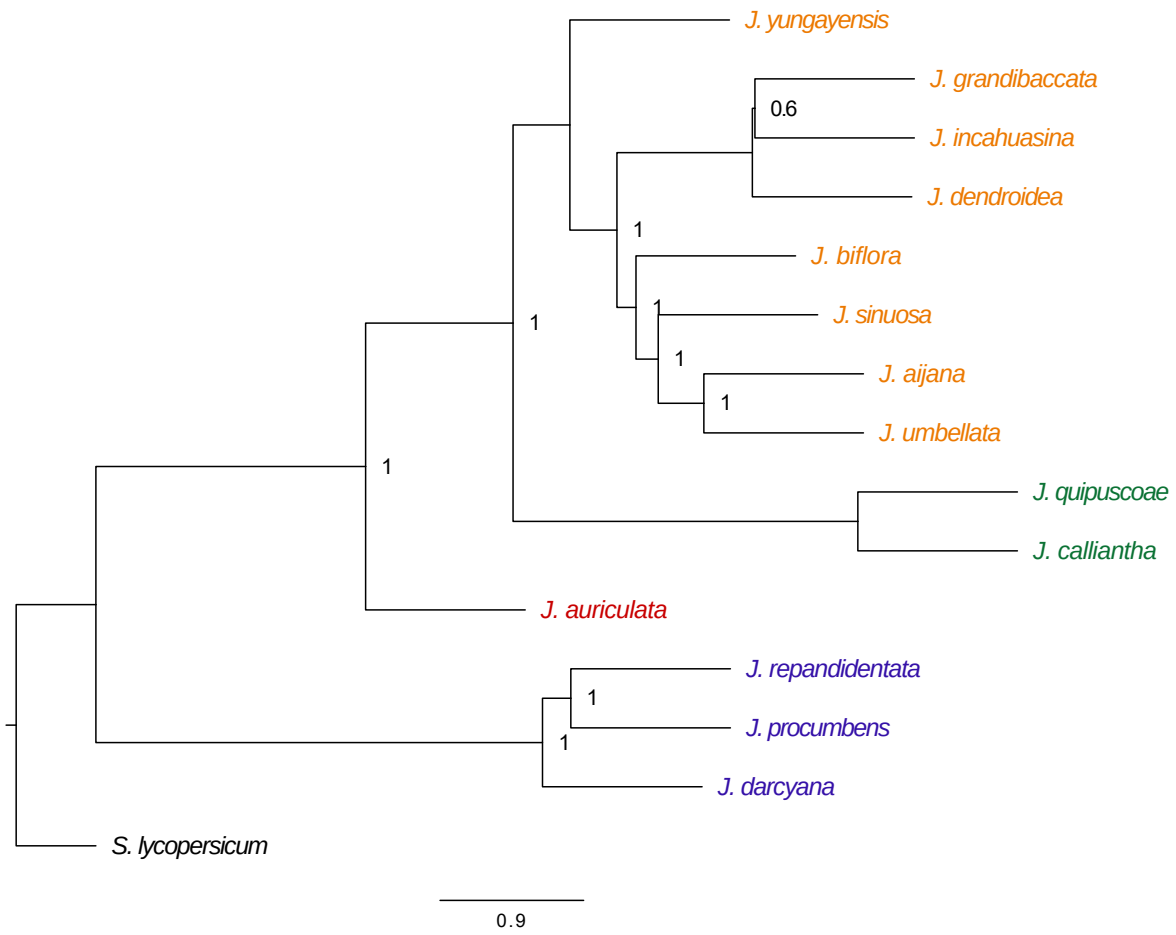

Figure S30 – ASTRAL tree for *Jaltomata*. Numbers at nodes are local posterior probabilities. Coloring of tip labels and outlined clades indicate fruit color and follow the convention of the original study. The ASTRAL tree matches the original authors' concatenated ML tree with the exception of the switched placement of *J. dendroidea* and *J. incahuasina*. This relationship has a low local posterior probability. Internal branch lengths are coalescent units $\left(\frac{2\Delta\tau}{\theta}\right)$, while terminal branch lengths are not estimated by ASTRAL and arbitrarily set to 1. Note that ASTRAL estimates unrooted trees and the species tree is rooted with *S. lycopersicum*, so the branch subtending the *Jaltomata* MRCA is set arbitrarily to 0.5.

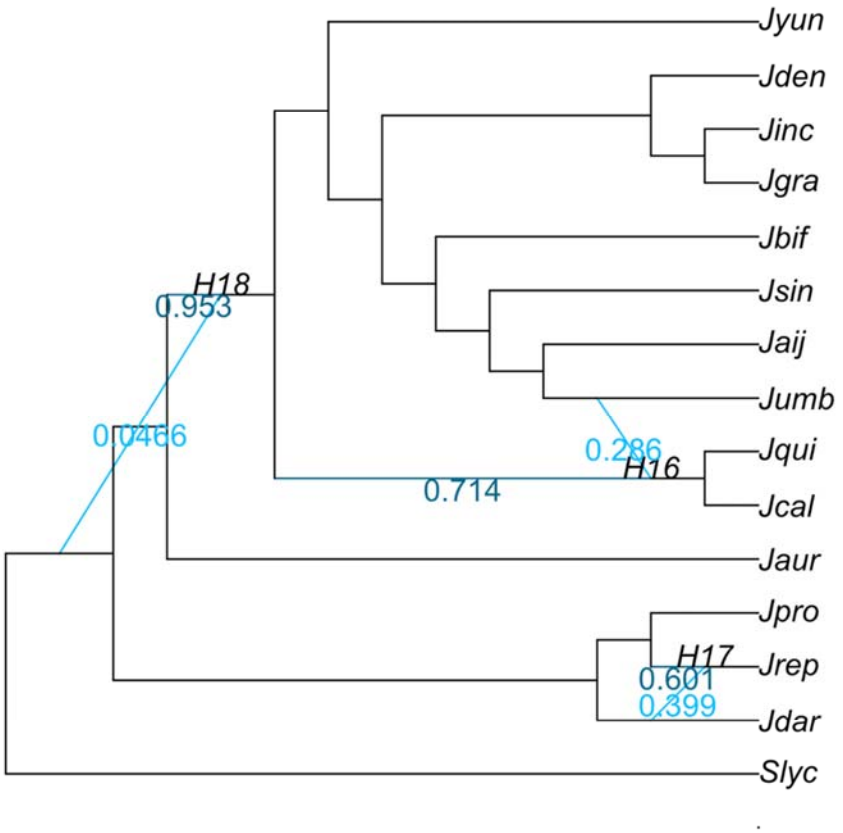

Figure S31 – Estimated Phylogenetic Network from PhyloNetworks for *Jaltomata*. The major species tree is identical to the ASTRAL tree. The admixture proportions are shown next to the parental branches at each hybridization node, with the black branches (corresponding to the greater admixture probability) representing the major species tree. The inferred introgression events (but not the introgression probabilities) were assumed in the MSci model to estimate divergence times with BPP.

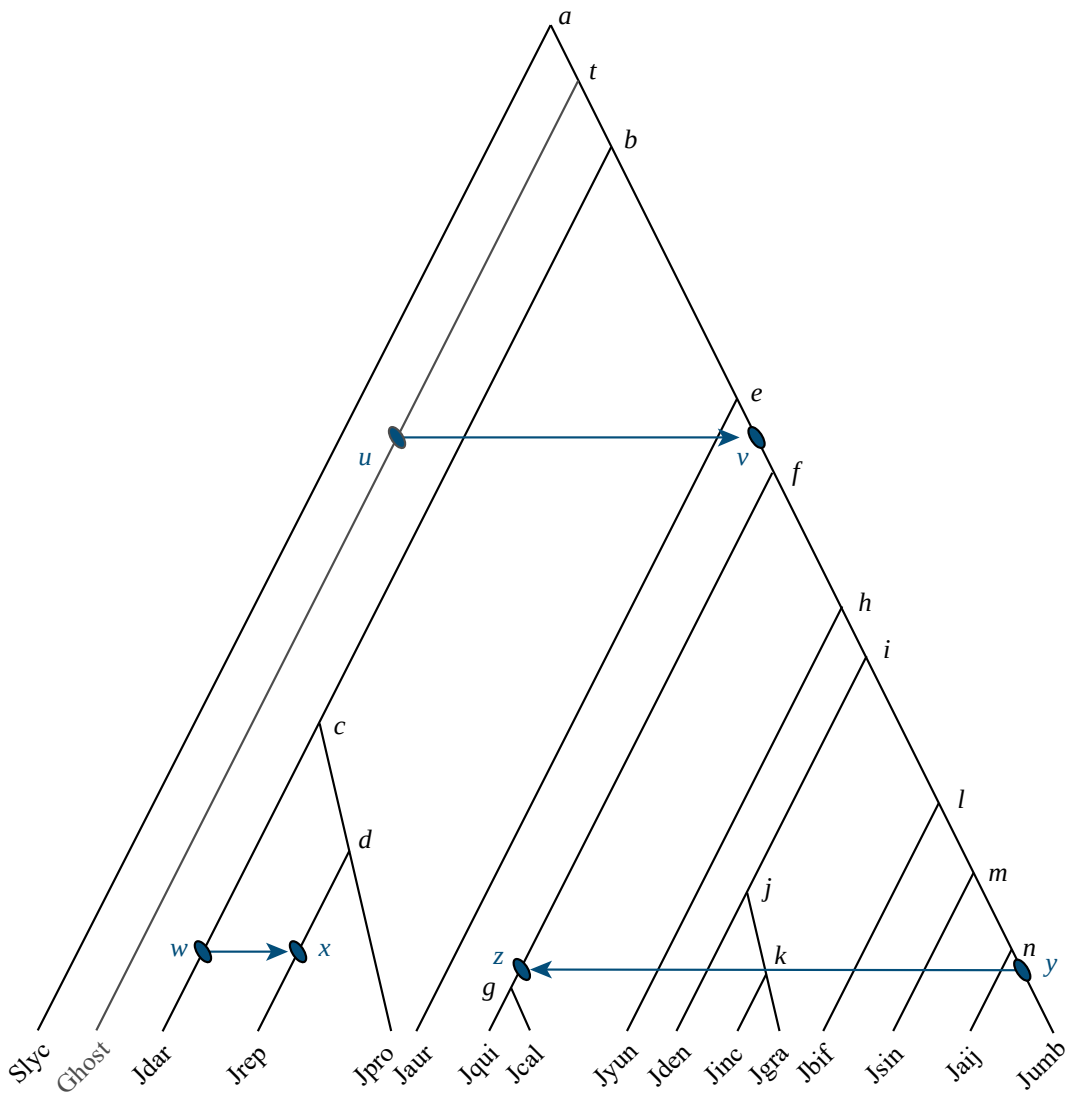

Figure S32 – Species tree with Node Labels Assumed in BPP Analyses. Blue edges represent introgression events inferred using PhyloNetworks and assumed in the BPP analysis under the MSci model to estimate divergence times. Node labels are used to identify parameters. A ghost lineage was included to model introgression from an unsampled lineage sister to *Jaltomata* into the common ancestor of the green- and orange-fruited clades.

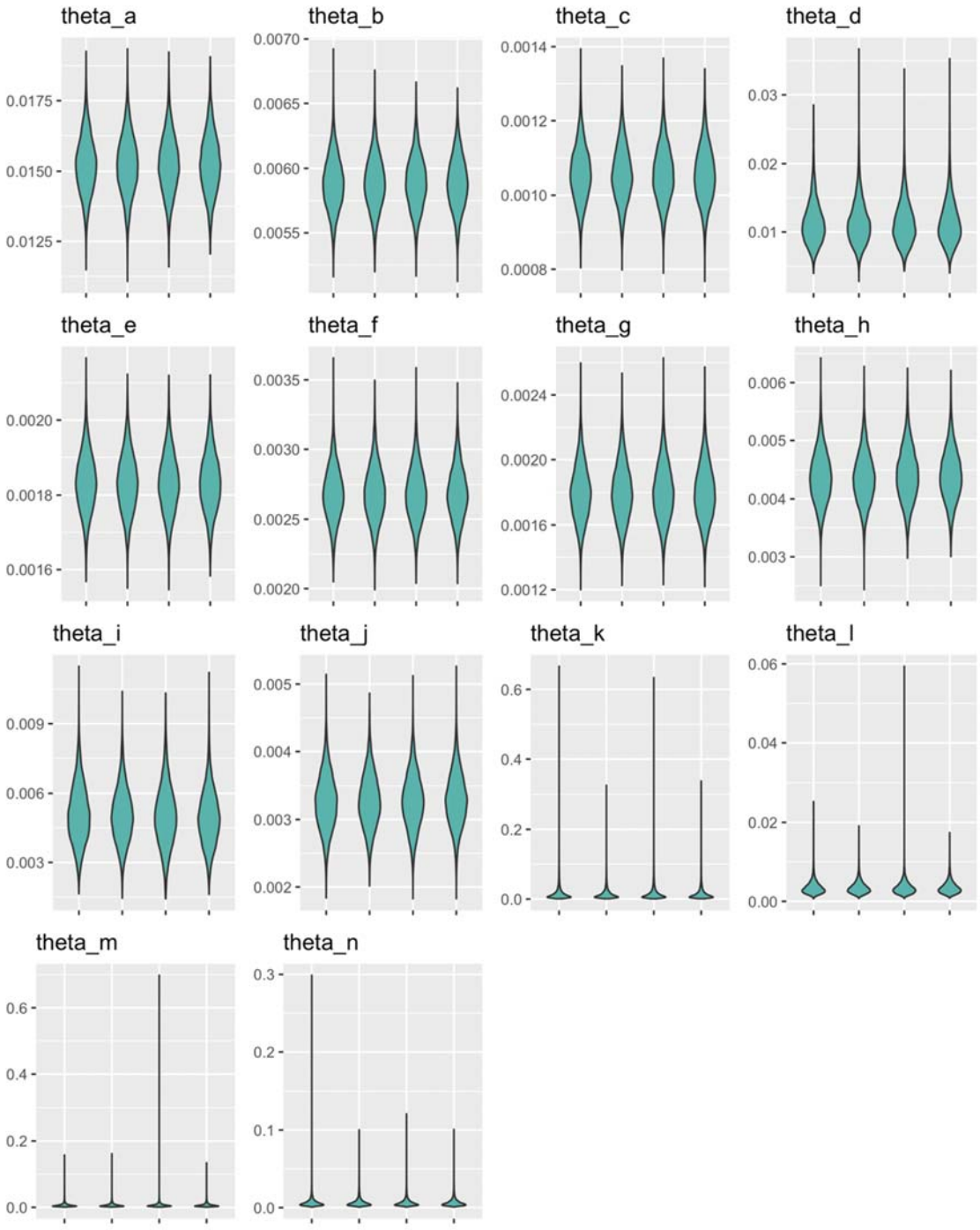

Figure S33 – Posterior Estimates of Population Sizes under the MSci in four Replicate runs for *Jaltomata*.

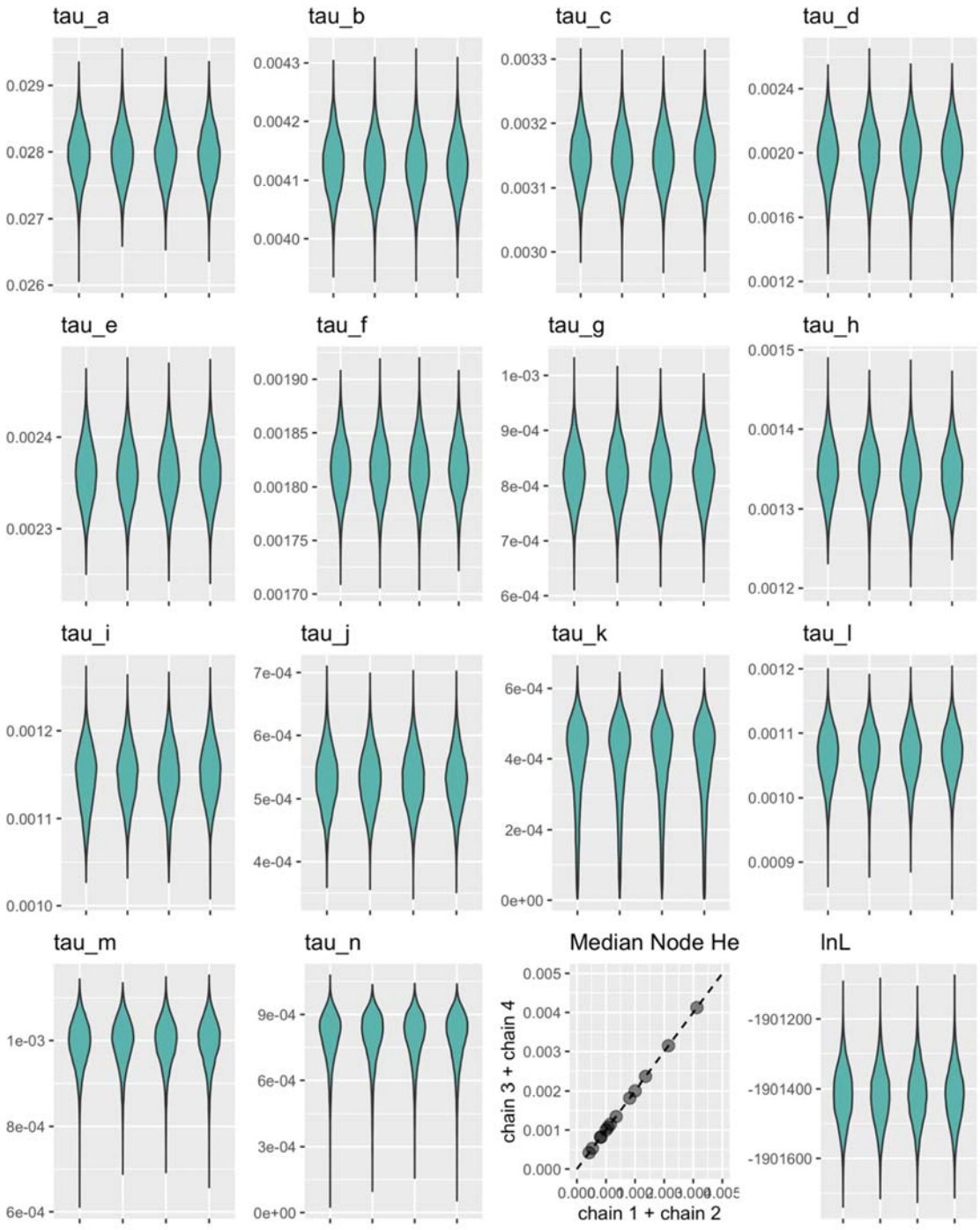

Figure S34 – Posterior Estimates of Divergence Times under the MSC in Four Replicate Runs for *Jaltomata*. The scatterplot of median node ages shows consistency between runs. The root node is excluded from the scatterplot to improve visibility.

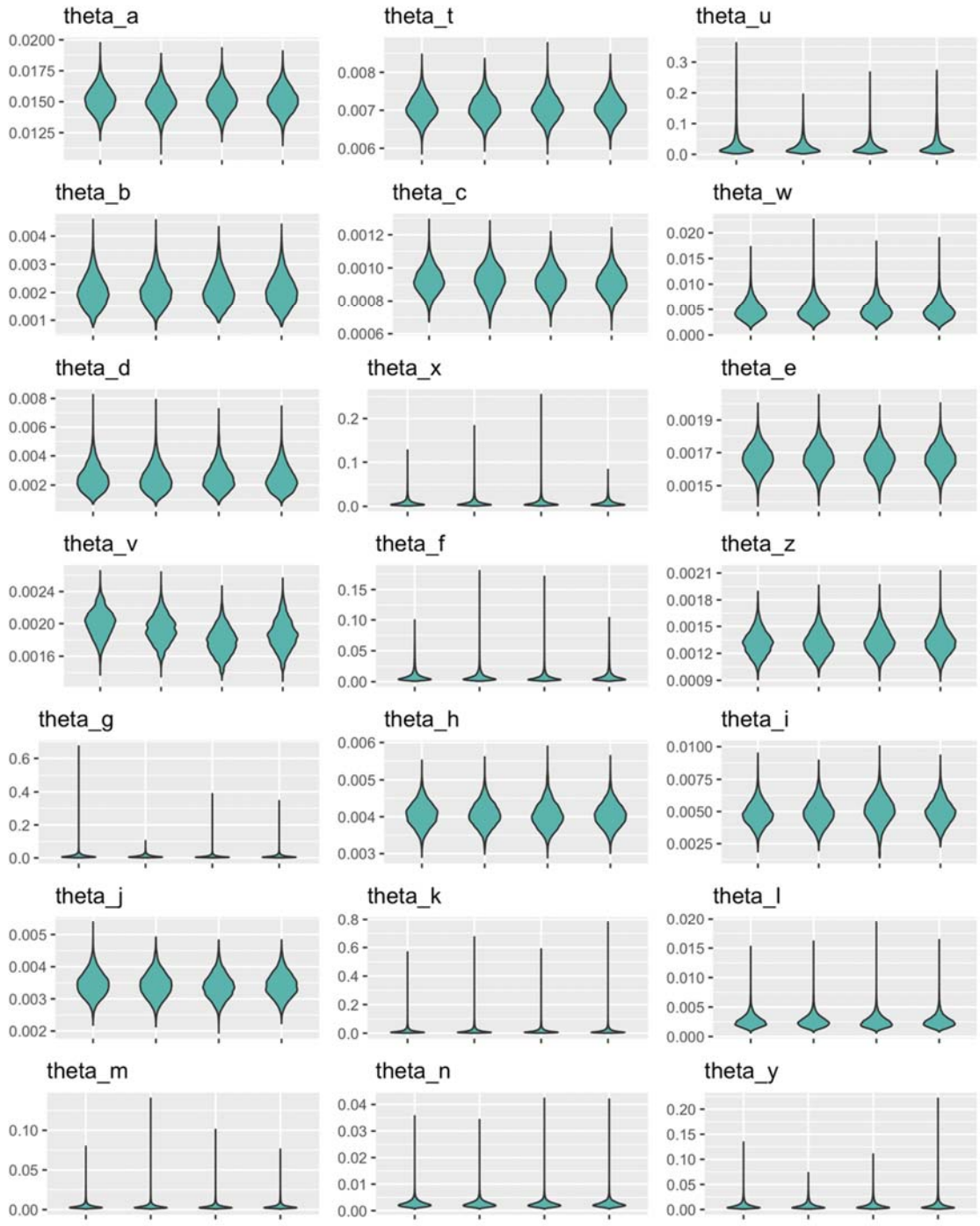

Figure S35 – Posterior Estimates of Population Sizes for *Jaltomata* under the MSci model in Four Replicate Runs.

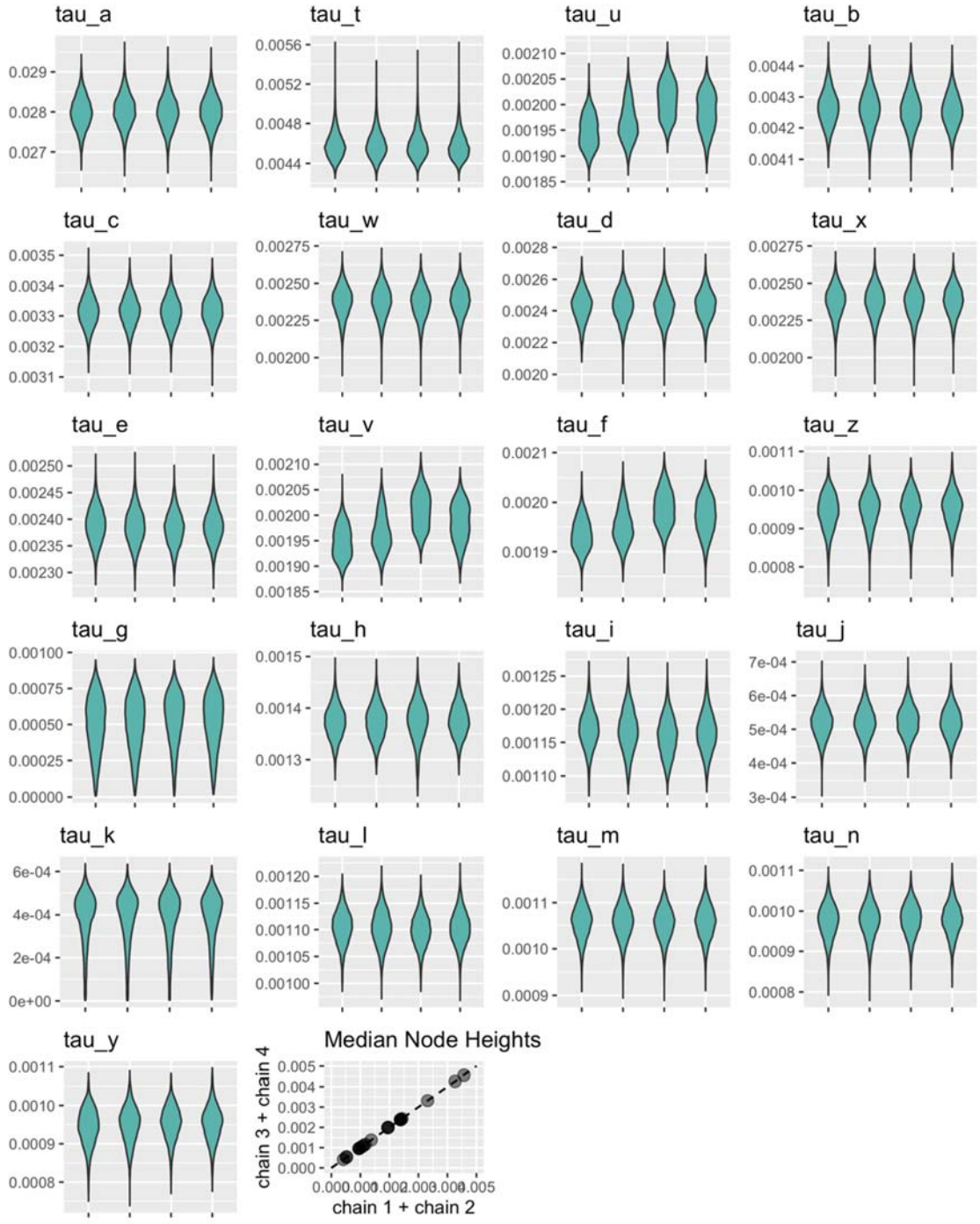

Figure S36 – Posterior Estimates of Divergence Times for *Jaltomata* under the MSci in Four Replicate Runs. The scatterplot of median node ages (τ) indicates convergence. The root node is excluded from the scatterplot to improve visibility.

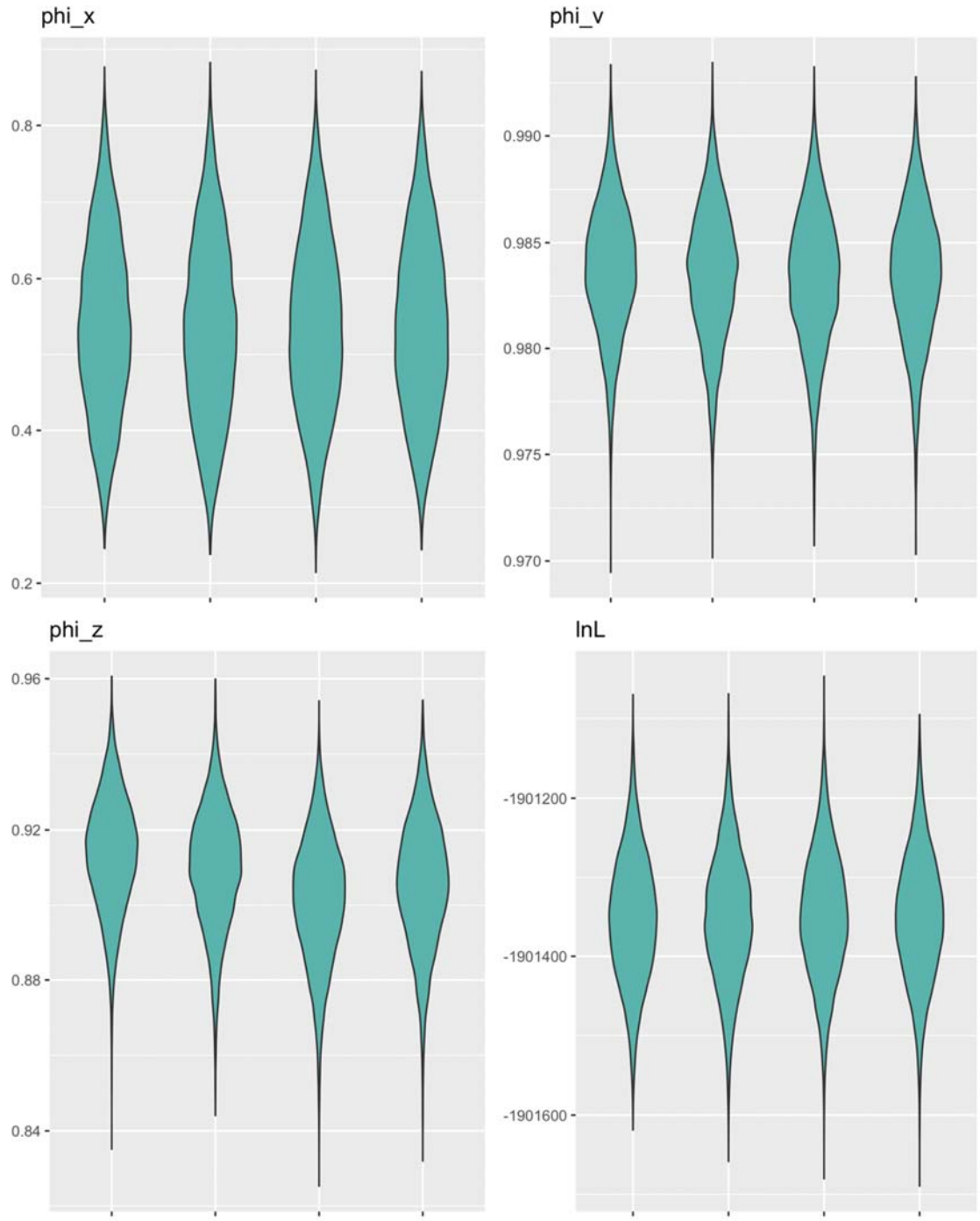

Figure S37 – Posterior Estimates of Introgression Probabilities for *Jaltomata* under the MSci model in Four Replicate Runs.

Supplementary Tables

	MSC				MSci			
Node Label	**Node Number**	**Mean**	**Lower 95% HPD**	**Upper 95% HPD**	**Node Number**	**Mean**	**Lower 95% HPD**	**Upper 95% HPD**
c	9	0.012883	0.012278	0.013466	9	0.01283	0.012229	0.013434
d	10	0.006206	0.005994	0.006428	10	0.00619 2	0.005981	0.00641
e	11	0.005875	0.005644	0.006095	11	0.00588 1	0.005648	0.006101
w	NA	NA	NA	NA	NA	NA	NA	NA
x	NA	NA	NA	NA	NA	NA	NA	NA
h	14	0.002275	0.002091	0.002454	15	0.00224 5	0.002056	0.002424
f	12	0.004029	0.003889	0.002454	12	0.00434 3	0.004145	0.004552
y	NA	NA	NA	NA	13	0.00246 1	0.001942	0.002877
z	NA	NA	NA	NA	16	0.00246 1	0.001942	0.002877
i	15	0.001074	0.000999	0.001148	17	0.00097 8	0.000896	0.00106
j	16	0.00106	0.000987	0.001138	18	0.00096 5	0.00088	0.001044
g	13	0.003994	0.00386	0.004129	14	0.00420 3	0.00404	0.004365

Table S1 – Divergence Time Estimates for *Adansonia*. Node labels correspond to nodes in Figure S25. Node numbers correspond to parameter order in BPP log files available through Dryad

Node Label	MSC				MSci			
	Node Number	Mean	Lower 95% HPD	Upper 95% HPD	Node Number	Mean	Lower 95% HPD	Upper 95% HPD
a	16	0.027963	0.027153	0.028723	16	0.028	0.027214	0.028789
t	NA	NA	NA	NA	17	0.004586	0.004341	0.004866
u	NA	NA	NA	NA	18	0.001977	0.001898	0.002057
b	17	0.004125	0.004032	0.00422	19	0.004259	0.004144	0.004369
c	18	0.003147	0.003063	0.003234	20	0.003315	0.003217	0.003412
w	NA	NA	NA	NA	21	0.002375	0.002156	0.002561
d	19	0.001995	0.001636	0.002332	22	0.002429	0.002233	0.00261
x	NA	NA	NA	NA	23	0.002375	0.002156	0.002561
e	20	0.002361	0.002297	0.002423	24	0.002386	0.002321	0.002449
v	NA	NA	NA	NA	25	0.001977	0.001898	0.002057
f	21	0.001817	0.001765	0.00187	26	0.001962	0.001885	0.002041
z	NA	NA	NA	NA	27	0.00095	0.000853	0.001032
g	22	0.000821	0.000721	0.000923	28	0.000517	0.000141	0.000828
h	23	0.001346	0.001276	0.001414	29	0.001374	0.001311	0.001442
i	24	0.00115	0.001079	0.001213	30	0.001165	0.001103	0.00122
j	25	0.000531	0.000438	0.000618	31	0.000521	0.000431	0.000608
k	26	0.000392	0.000119	0.000575	32	0.000387	0.00012	0.000566
l	27	0.001066	0.000981	0.00114	33	0.001103	0.001041	0.001164
m	28	0.000994	0.000879	0.001091	34	0.001057	0.000984	0.001123
n	29	0.000795	0.000563	0.000983	35	0.000976	0.000893	0.001053
y	NA	NA	NA	NA	36	0.00095	0.000853	0.001032

Table S2 – Divergence Time Estimates for *Jaltomata*. Node labels correspond to nodes in Figure S32. Node numbers correspond to parameter order in BPP log files available through Dryad