Supplementary Materials for “Phylogenetic Stochastic Mapping without Matrix Exponentiation”

by Jan Irvahn and Vladimir N. Minin

Appendix A

We present evidence supporting the claim that the stationary distribution of our new MCMC sampler is the posterior distribution, $p(\mathcal{V}, \mathcal{W}|\mathbf{Y})$. This posterior has many aspects that could be examined, but for simplicity we focus on univariate statistics: the amount of time spent in each state and the number of transitions between each pair of states.

We compare the results of five different implementations. The first is a sampler implemented in the diversitree package [FitzJohn, 2012], labeled diversitree or DIV. The second is our version of the same method, labeled EXP. The third is the same method that only exponentiates the rate matrix once, labeled EXP ONCE or ONCE. The fourth is our new method, labeled MCMC. The fifth is a sparse version of our new method, labeled SPARSE or SPA.

We present results for four regimes, two different sizes of state spaces, and two different sets of transition rates. The smaller state space has 4 states and the larger has 20. The lower transition rates correspond to 2 expected transitions per tree and the higher transition rates correspond to 20 expected transitions per tree. In an effort to reduce the number of plots we focus only on states that were observed at the tips of the tree. All four simulated trees had 20 tips.

Our first example used the smaller state size, 4, with the smaller number of expected transitions per tree, 2. A random simulation resulted in two unique tip states, states 1 and 3. For each method we produced 100,000 state history samples. Figure A-1 contains plots of four univariate statistics pulled from the posterior distributions. All five implementations produced the same results.

Our second example used the smaller state size, 4, with the larger number of expected transitions per tree, 20. A random simulation resulted in three unique tip states, states 1, 2, and 4. Figure A-2 contains four boxplots pulled from the posterior distributions. Figure

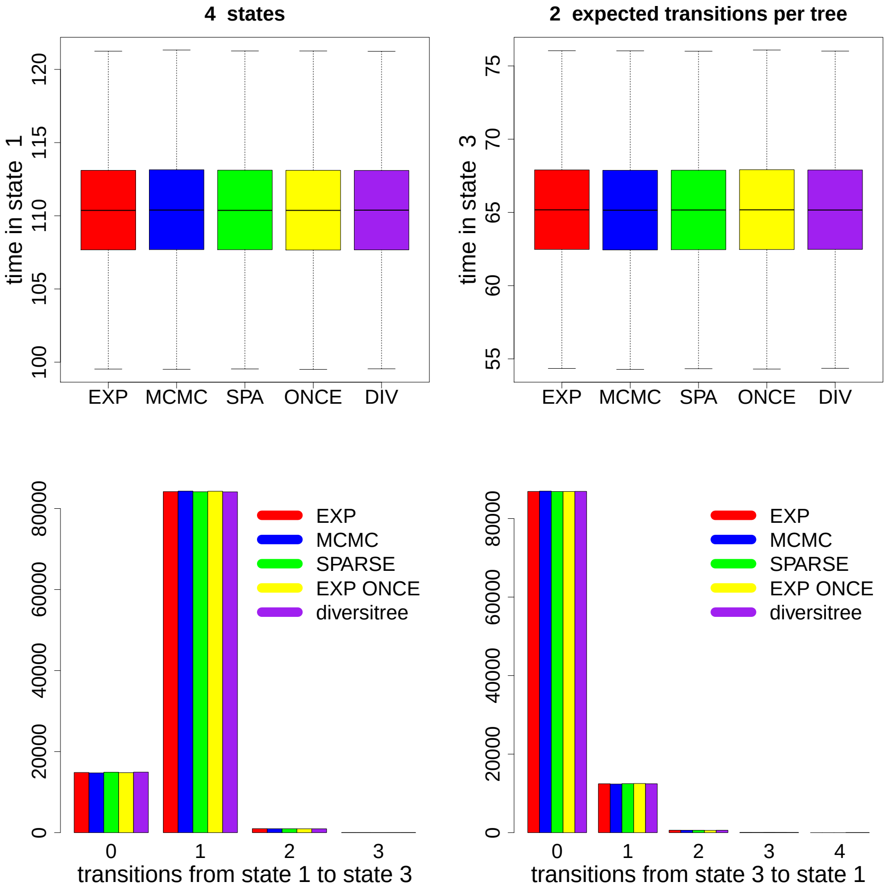

Figure A-1: Univariate summaries for five implementations of state history sampling of a 20 tip tree. There were 4 states and 2 expected transitions per tree. The top plots contain boxplots illustrating the distribution of the amount of time spent in state 1 and state 3. Outliers were not included though all five implementations showed the same outlier behavior. The bottom plots contain histograms illustrating the posterior distribution of the number of transitions between state 1 and state 3.

A-3 contains six histograms pulled from the posterior distributions.

Our third example used the larger state size, 20, with the smaller number of expected transitions per tree, 2. A random simulation resulted in two unique tip states, states 1 and 5. Figure A-4 contains plots of four univariate statistics pulled from the posterior distributions

Our fourth example used the larger state size, 20, with the larger number of expected transitions per tree, 20. A random simulation resulted in six unique tip states, states 1, 4, 8, 10, 12, and 15. Figure A-5 contains six boxplots pulled from the posterior distributions.

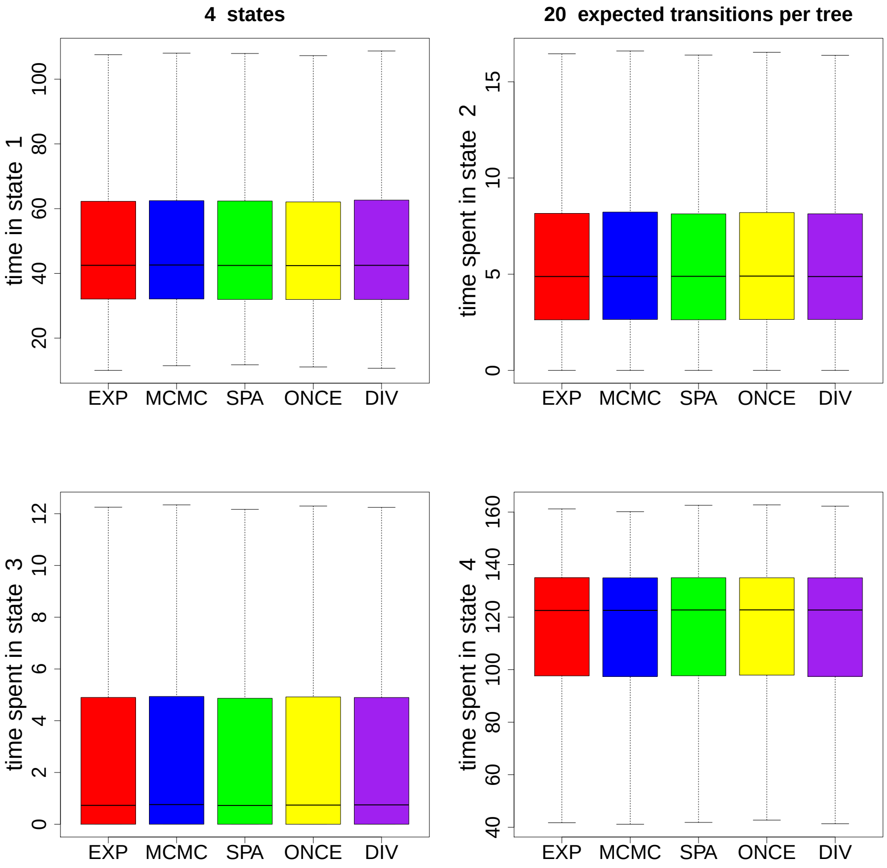

Figure A-2: Boxplots illustrating the posterior distribution of the amount of time spent in each state. Outliers were not included though all five implementations showed the same outlier behavior. There were 4 states and 20 expected transitions per tree.

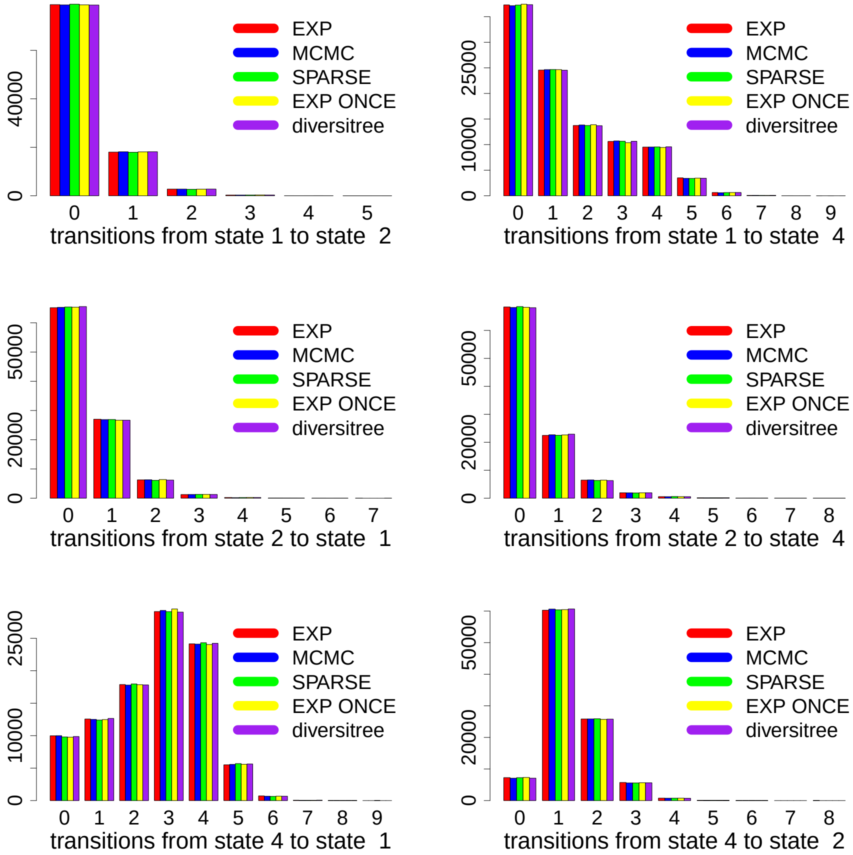

Figure A-3: Histograms illustrating the posterior distribution of the number of transitions between states 1, 2, and 4. There were 4 states and 20 expected transitions per tree.

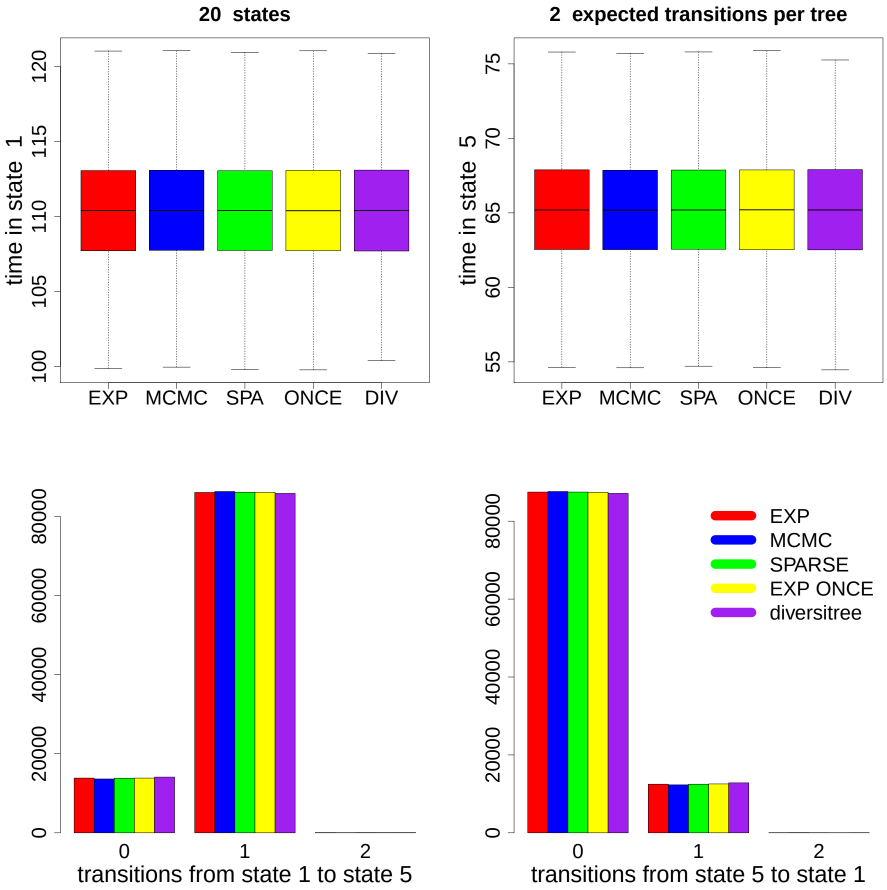

Figure A-4: Univariate summaries for 5 implementations of state history sampling of a 20 tip tree. There were 20 states and 2 expected transitions per tree. The top plots contain boxplots illustrating the posterior distribution of the amount of time spent in state 1 and state 5. Outliers were not included though all five implementations showed the same outlier behavior. The bottom plots contain histograms illustrating the posterior distribution of the number of transitions between state 1 and state 5.

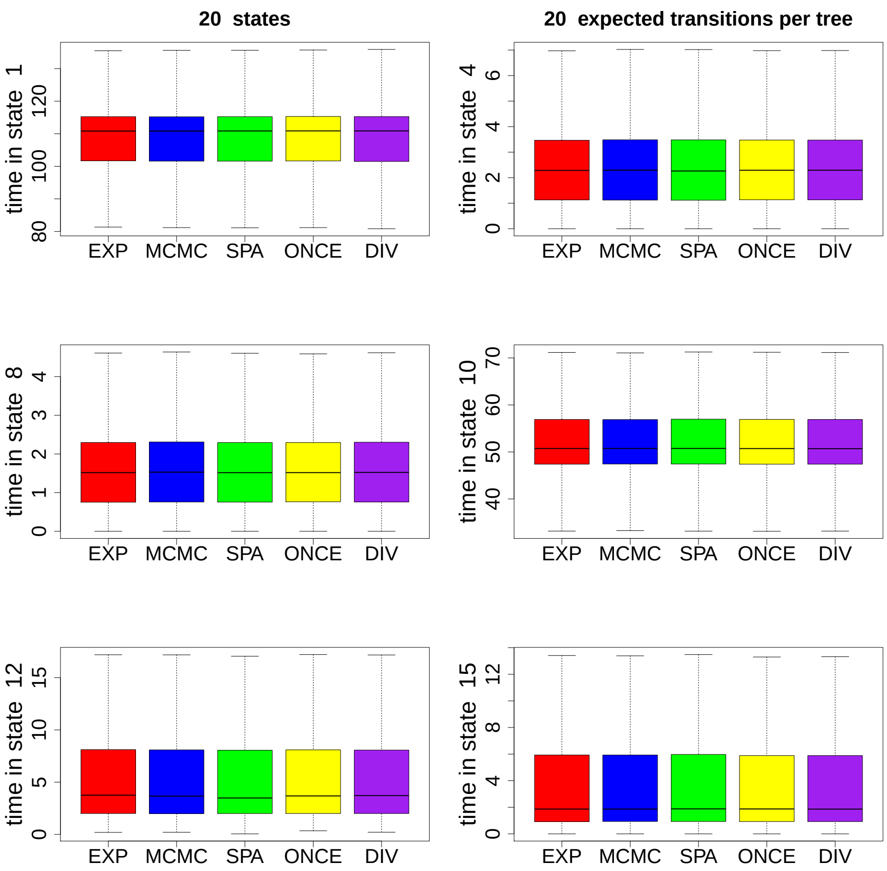

Figure A-5: Boxplots illustrating the posterior distribution of the amount of time spent in each tip state. Outliers were not included though all five implementations showed the same outlier behavior. There were 20 states and 20 expected transitions per tree.

Figure A-6 contains six histograms pulled from the posterior distributions.

Appendix B

One state space of interest in molecular evolution is the amino acid state space. Jones et al. [1992] proposed a rate matrix for an amino acid CTMC substitution model, called JTT. Figure B-1 contains timing results for this rate matrix and a tree with 40 tips. When our MCMC approach used an appropriately tuned value of Ω we saw slightly faster running times as compared to the matrix exponentiation approach. We examined other scenarios for the JTT rate matrix in which we saw faster running times with the matrix exponentiation

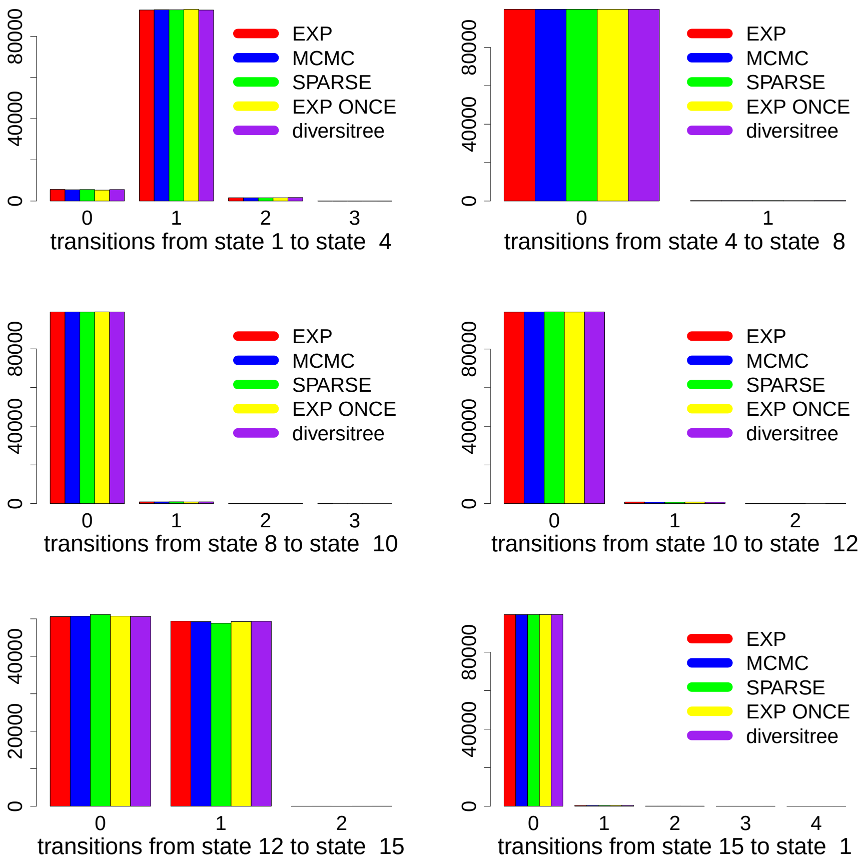

Figure A-6: Histograms illustrating the posterior distribution of the number of transitions between a subset of the tip states. There were 20 states and 20 expected transitions per tree.

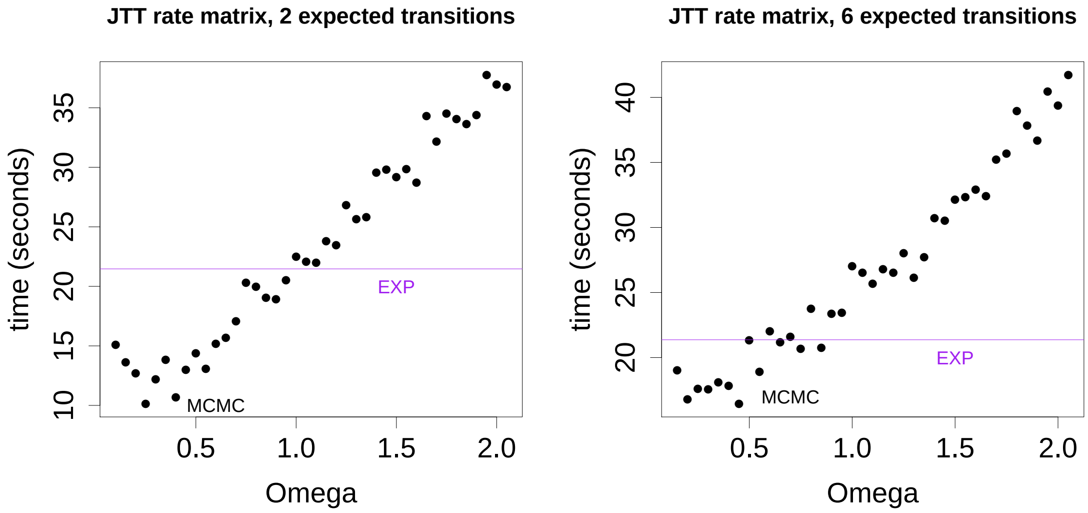

Figure B-1: Time to obtain 10,000 effective samples as a function of the dominating Poisson process rate, Ω, for the JTT amino acid rate matrix as found in the phylosim R package. Results for our MCMC sampler are shown in black. Timing results for the matrix exponentiation method are represented by a purple horizontal line because the matrix exponentiation result does not vary as a function of Ω. The randomly generated tree had 40 tips. The JTT rate matrix was scaled to produce 2 expected transtions in the left hand plot. The JTT rate matrix was scaled to produce 6 expected transtions in the right hand plot.

approach.

Appendix C

Our MCMC sampler seems to converge to stationarity quickly. Figure C-1 shows two convergence plots, one for fast evolution and one for slow evolution. In both cases we started the chain with an augmented substitution history containing one transition in the middle of each branch leading to a tip whose state was different from an arbitrarily chosen root state. In the case of slow evolution this substitution history was a poor starting point but the log likelihood of the chain appeared to achieve stationarity quickly. In both cases, the tree had 50 tips and the size of the state space was 10.

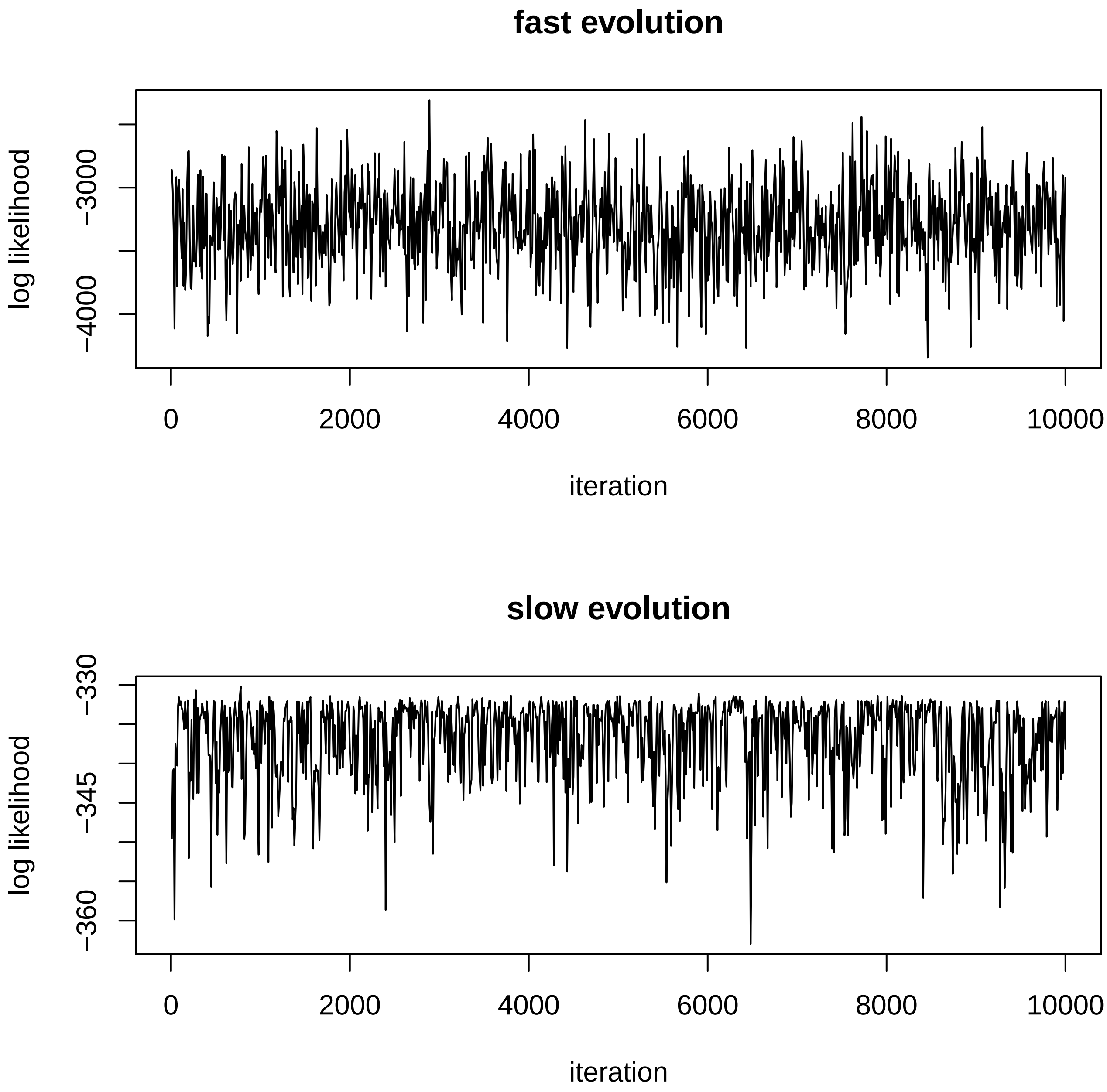

Figure C-1: MCMC trace plots. We show the log density of substitution histories for two MCMC chains at every tenth iteration. The top plot shows results for a trait that evolved quickly (with 6 expected substitutions). The bottom plot shows results for a trait that evolved slowly (with 2 expected substitutions). In both cases, the tree had 50 tips and the size of the state space was 10.

Appendix D

Our MCMC method scales well with the size of the state space even when state space sizes exceed 100. Figure D-1 shows timing results for state space sizes going out to 300. We show results for our MCMC method and a sparse version of our MCMC method using tridiagonal rate matrices.

References

Richard G FitzJohn. Diversitree: comparative phylogenetic analyses of diversification in R. Methods in Ecology and Evolution, 3(6):1084–1092, 2012.

David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. Computer applications in the biosciences: CABIOS, 8(3):275–282, 1992.

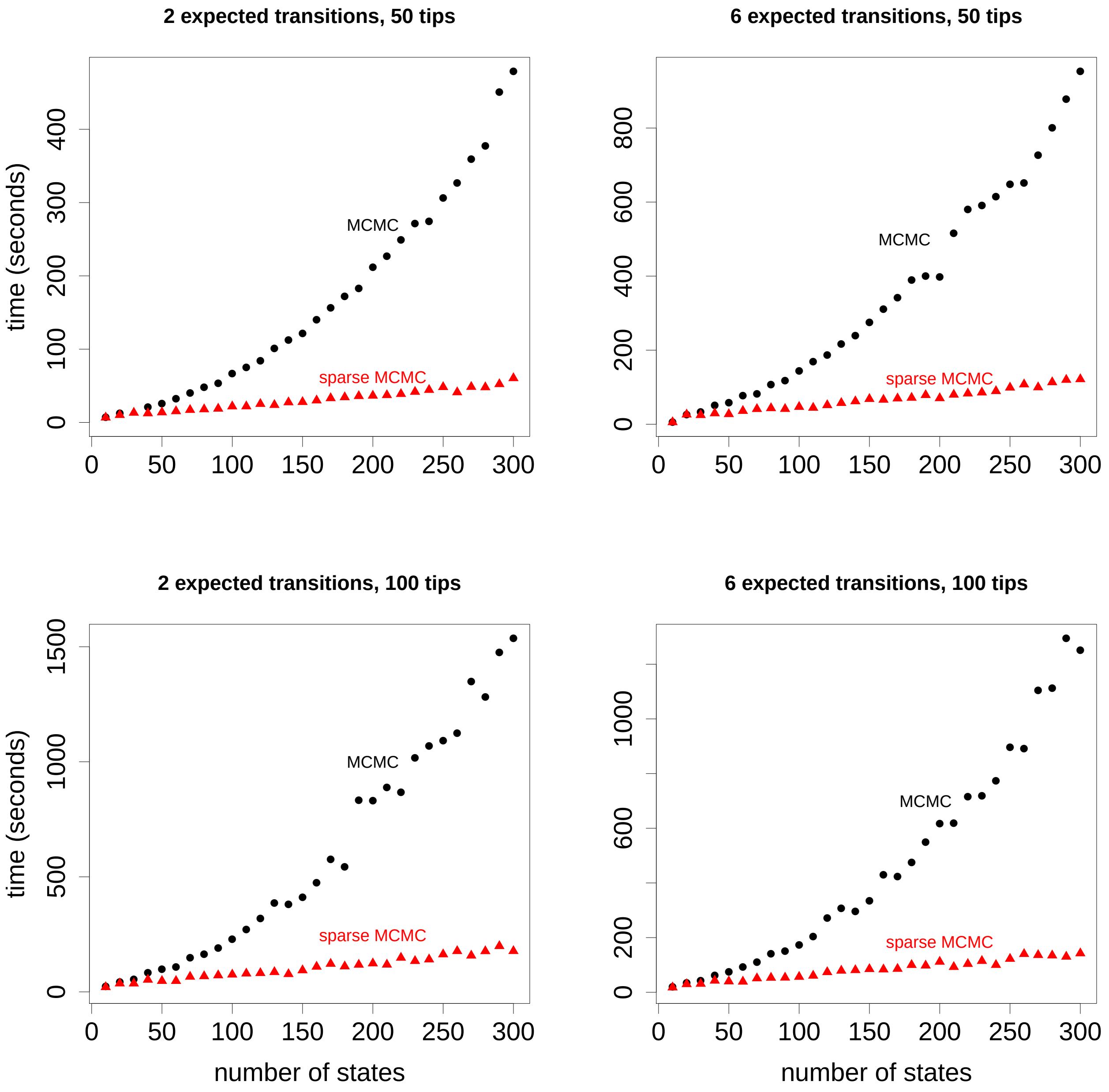

Figure D-1: State space effect for a tridiagonal rate matrix. All four plots show the amount of time required to obtain 10,000 effective samples as a function of the size of the state space for two methods, our MCMC sampler in black circles and a sparse version of our MCMC sampler in red triangles. The two plots in the top row show results for a randomly generated tree with 50 tips. The two plots in the bottom row show results for a randomly generated tree with 100 tips. The two plots in the left column show results for a rate matrix that was scaled to produce 2 expected transitions while the two plots in the right column show results for a rate matrix that was scaled to produce 6 expected transitions.