1 SUPPLEMENTARY MATERIALS

1.1 Alternative Models of Rate Heterogeneity

The rate heterogeneity simulated for the data in Section 4.1 allows for specific regions of the alignment that are under different evolutionary pressures, expressed with a step-function like change. However, in general the variation in evolutionary rates may not follow such a sharp shift in distribution. In order to test the robustness of the STHMM to other types of rate variation, we simulated data under a Gamma model of site-specific rate variation. This was done using the Seq-Gen program of Rambaut and Grassly (1997). The same topologies were used as for the previous simulation, shown in Figure 3 of the main paper. Each site is then assigned a 'relative rate of evolution' by Seq-Gen which comes from a Gamma distribution with a shape parameter set by the user. The distribution is then scaled so that the rates have a mean of 1. For this simulation the shape parameter was set to 0.3 which allows for a large degree of site-specific rate heterogeneity.

Figure 1 shows the results from the STHMM for the Gamma distributed rate variation. We have successfully recovered the correct recombinant structure that the data was simulated under. The STHMM is now less certain about the locations of the topology breakpoints. The evolutionary rate has been captured in the parameter μ shown in the bottom panel. The frequent switches in the choice of value for μ reflect the continuously varying rate. Since our values of μ are restricted to the grid $[0.001, 0.003, 0.01, 0.03, 0.1, 0.3, 1.0, 3.0, 10.0, 100.0]$ the posterior probability is often shared between two or more states at each site since the true value of μ may not correspond precisely with on of the HMM states.

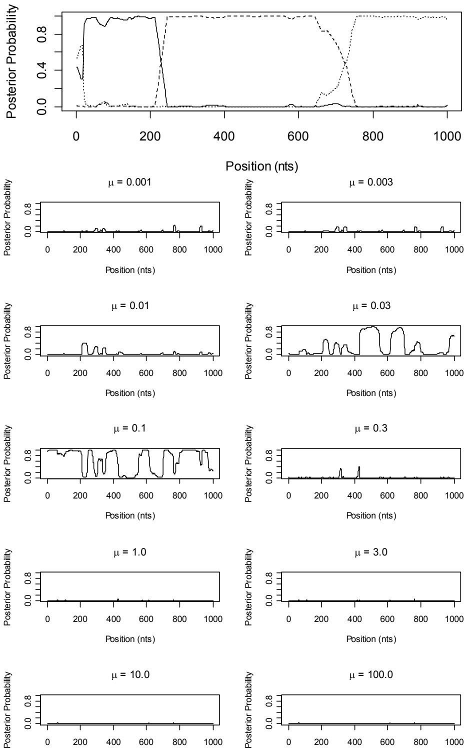

Fig. 1. The posterior probability for the topologies and rates inferred by the STHMM. We have recovered the true topolgies and the recombinant structure. The positions of the changes in topology are less precise than those with the step-function like rate variation. We have also inferred a continuous variation in evolutionary rate as we move along the alignment.

1.2 Effect of Substitution Rate on Recombination Detection

Our ability to detect recombination depends crucially on the amount of variation we see in the data. In the extreme case of no variation, the data will be just as likely under any topology and we will not be able to infer recombination. In order to investigate how sensitive the STHMM is to the substitution rate we have simulated further data under differing substitution rates. In order to mimic the substitution rate, the branch lengths of the topologies used to simulate the data using Seq-Gen (Rambaut and Grassly, 1997) were scaled. This has the effect of changing the expected number of substitutions per site. 50 replicates were run for several different values of this scaling parameter and in each case the inferred topologies were recorded. Table 1 compares the success rates of topology recovery for the different simulations.

The results for the scaling factor of 0.01 are not surprising, since there was very little diversity between the taxa and so many of the trees had equal likelihood and the STHMM was unable to distinguish between them and infer recombination. For the scaling factor of 0.1 the STHMM has inferred recombination in the majority of cases but since there is still little diversity amongst the taxa it was unable to recover the true topologies in some cases, particularly topology one which is the shortest section. When the scaling factor is 0.5 (i. e. branch lengths of 0.05) the STHMM is able to infer recombination and also the correct topologies.

Table 1. Comparison of the topology recovery rate and the breakpoint location for several different values of the branch length scaling parameter. For each value, 50 replicates were generated. The maximum posterior probability topology for each region was considered and the number of times the correct topology was used in each region was recorded. The standard errors for the breakpoints are shown in brackets.

	Scaling Parameter		
	0.01	0.1	0.5
Topology 1	0.08	0.46	1.0
Topology 2	0.16	0.98	1.0
Topology 3	0.04	0.86	1.0
Breakpoint 1	-	209.12 (9.36)	201.02 (1.69)
Breakpoint 2	-	707.74 (5.24)	706.36 (1.51)

1.3 Convergence of the MCMC Scheme

Since we are unable to calculate exact posterior probabilities we rely on a MCMC scheme to simulate draws from the posterior. We

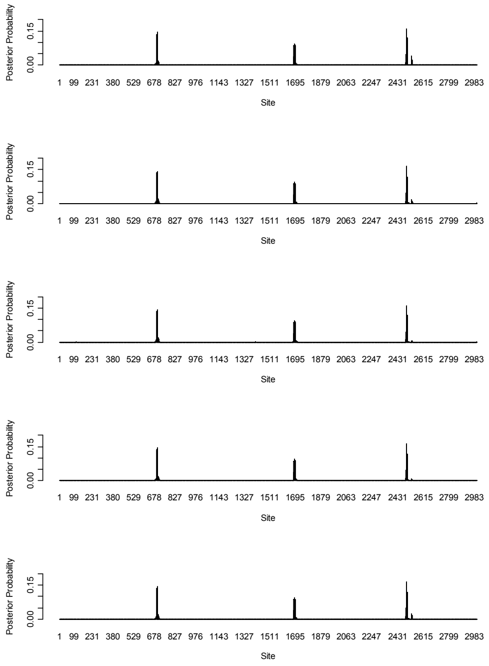

Fig. 2. The posterior probabilities of a breakpoint at each site in the alignment for five separate runs of the algorithm. The spikes in probability show that the algorithm is inferring breakpoints in these regions.

would therefore like to know if our algorithm is converging to the correct target distribution. However, it is not possible to verify that the Markov chain has converged and we are indeed drawing samples from the posterior and only corroboratory diagnostics exist. To investigate the convergence of our algorithm, we have run the method five times on the same dataset (the 15 taxa dataset discussed in the main paper) and in each case estimated the posterior probability of a breakpoint at each site. This was calculated by simply counting the number of times a breakpoint was sampling from the last 50,000 iterations of the algorithm and dividing by the number of iterations. These probabilities are shown in Figure 2.

As we can see, the algorithm consistently recovers the correct breakpoints. This provides us with some evidence that the algorithm is converging and we are indeed drawing samples from the posterior distribution.

REFERENCES

Rambaut, A. and Grassly, N. C. (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees, *Comput. Appl. Biosci.*, **13**, 235-238.