

# HMM based genomic investigation of the divergence of two closely related species with application to the Bornean and Sumatran orangutan — Supplemental Material

Thomas Mailund<sup>1,\*</sup>, Julien Dutheil<sup>1</sup>, Asger Hobolth<sup>1,2</sup>, Gerton Lunter<sup>3</sup>, Mikkel Heide Schierup<sup>1,4</sup>

**1** Bioinformatics Research Center, Aarhus University, Aarhus, Denmark

**2** Department of Mathematical Sciences, Aarhus University, Aarhus, Denmark

**3** The Wellcome Trust Centre for Human Genetics, Oxford, United Kingdom

**4** Department of Biology, Aarhus University, Aarhus, Denmark

\* E-mail: mailund@birc.au.dk

## 1 Model validation

The power and consistency of the model can be evaluated through simulations. The HMM is designed to approximate the coalescent with recombination process. Thus, we simulate data using coalescent with recombination and then compare inferred parameters from the HMM model with the values used in the simulations.

### 1.1 The Markov approximation

The CTMC approach only approximately captures how coalescence times are distributed in time, along the sequences, since it only considers two nucleotides and then assumes that the process is Markovian along the sequence.

Previously, in Dutheil *et al.* we showed that the coalescent process leading to segments in incomplete lineage sorting approximates a geometric distribution, justifying a Markov approximation used in the CoalHMMs. There, we showed that the distribution of nucleotides in incomplete lineage sorting follows a geometric distribution with the same parameter as inferred in the CoalHMM.

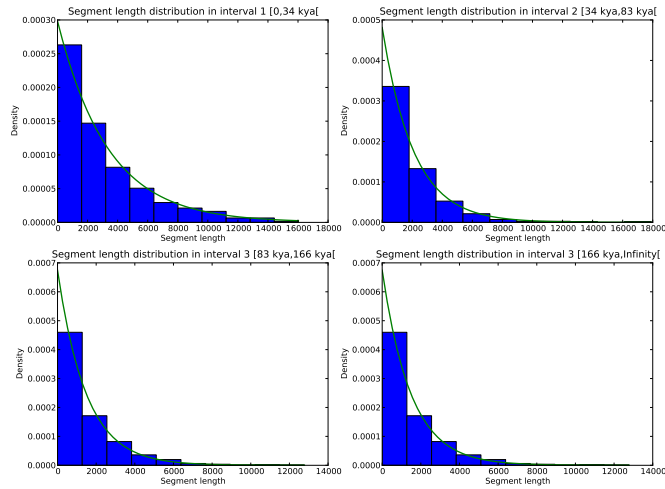
For the new model, we needed to test how the changes in coalescence times – the distribution of segments finding their most recent common ancestor within a given time interval – is distributed, and how this distribution compares to the geometric distribution predicted by the Markov approximation.

To test this we simulated 50 replicates of the coalescence with recombination process, grouped segments with MRCA in the same time interval and calculated the length of contiguous segments within each time interval. The simulations were with parameters close to the inferred parameters for the orangutan sub-species: 30,000 for the effective population size, 1.5 cM/MB for the recombination rate and 300,000 years for the sub-speciation.

As shown in Fig. 1 the match between the empirical distribution of segment lengths and the predicted geometric distribution of the segment lengths from the CoalHMM are very close, leading us believe that the approximation we make is accurate enough for the inference we do.

### 1.2 Transition probabilities

To check that our CTMC calculations correctly computes the transition probabilities compared to coalescent simulation, we compared the computed with the observed transitions from one time interval to another. Figure 2 shows the results for a model with four coalescence time intervals. The box plots show the distribution of transition probabilities from 25 simulations and the blue dots show the mean transition probabilities from the simulations. The red dots show the transition probabilities derived by the CTMC calculations. In general, the fit between the empirical distribution and the CTMC calculation is very good.



**Figure 1. Distribution of segment lengths.** The plot compares the empirical distribution of segment lengths in the coalescent-with-recombination process in each time interval (blue histogram) with the expected geometric distribution of segment lengths using the Markov model approximation (green line).

### 1.3 Likelihood surface

The parameters of the model we estimate are 1) the population divergence time / sub-speciation time, 2) the effective population size of the ancestral population, and 3) the recombination rate. We simulated a 500kbp alignment with the following parameters: 1) a divergence time of 2 Mya, 2) an effective population size for the two sub-species of 10,000 and an ancestral effective population size of 100,000, and 3) a recombination rate of 1.5 cM/Mb.

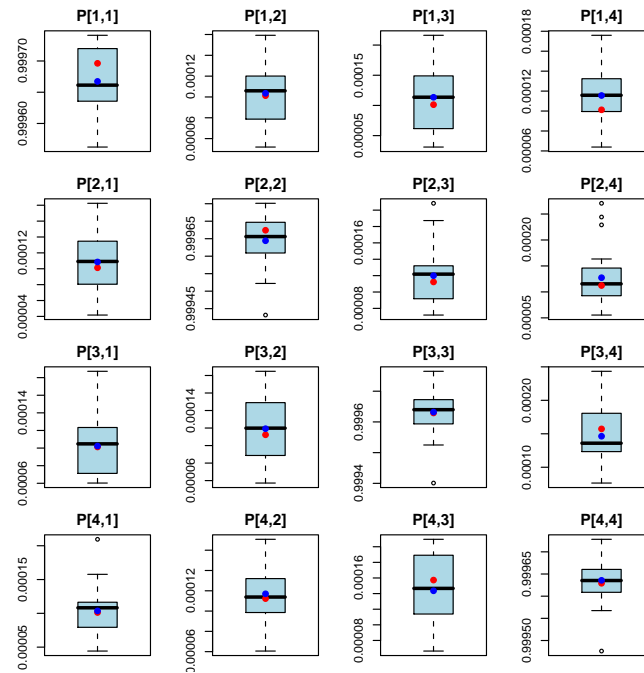
We then examined the likelihood of the data when running the model with 10 states, see Fig. 3. For the sub-speciation time and the ancestral effective population size, the maximum likelihood is found at the true parameter value, while for the recombination rate the maximum is slightly lower. The likelihood curve is flat for the recombination rate, and to a lesser degree for the divergence time, so we expect our estimates of these parameters to have a large variance, a problem that can hopefully be alleviated by the very large data set provided by a whole genome analysis.

### 1.4 Testing model assumptions

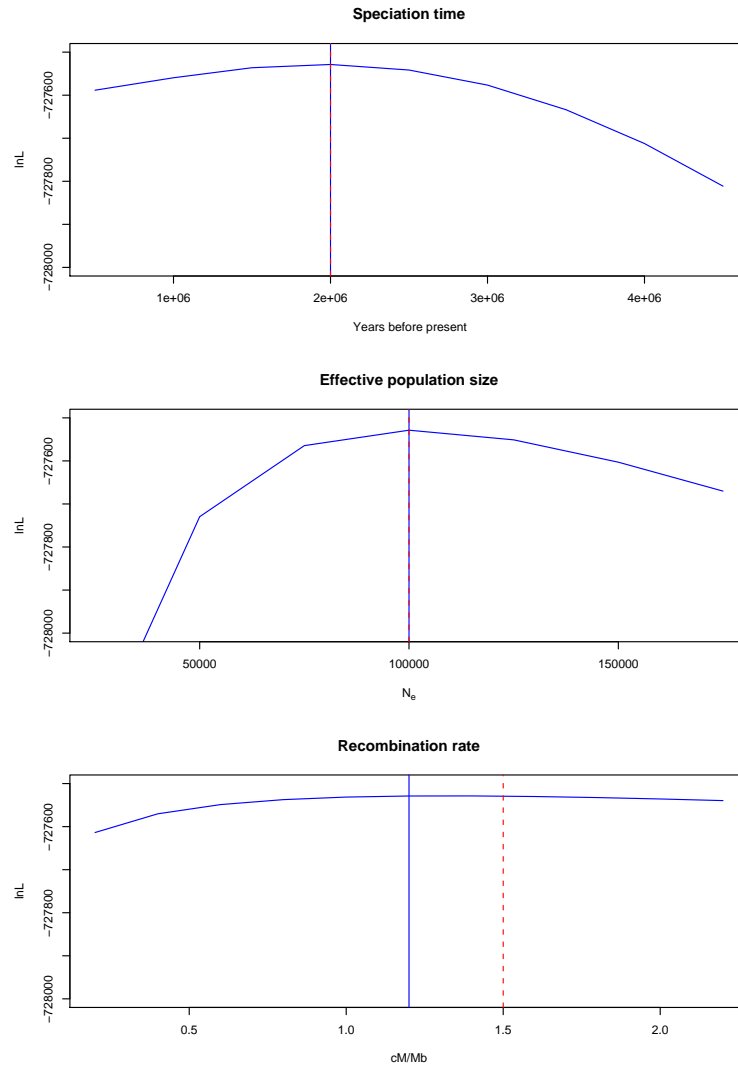
The hidden Markov model presented in the paper is an approximation to the coalescent process with recombination and to the true underlying process giving rise to the genomic data we analyze. While testing the consequences of the approximations to the true process is in nature difficult as we do not *know* all the factors underlying the true process, we can test the impact of the approximations to the coalescent process through simulations.

#### 1.4.1 Comparing the hidden Markov model with the coalescent process

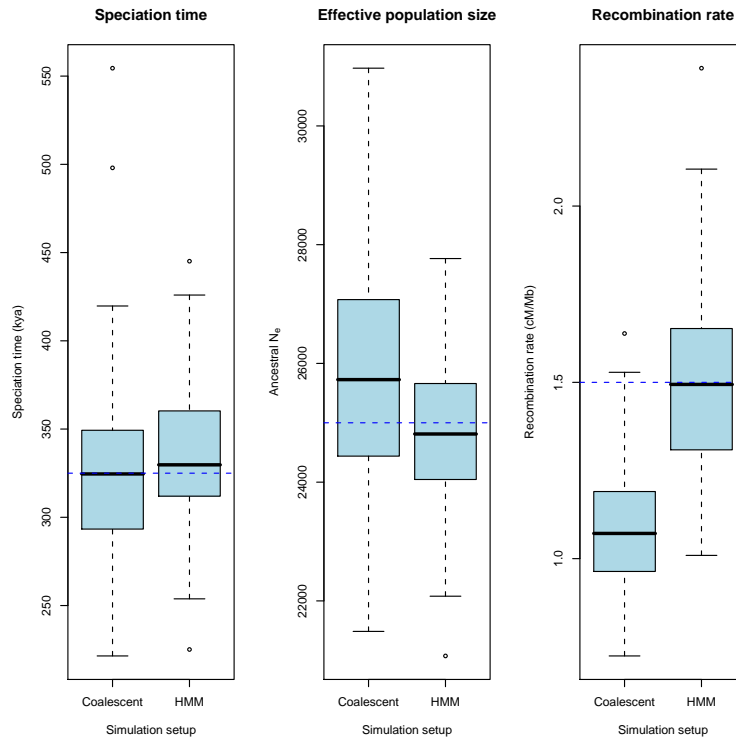
When simulating alignment data directly from the hidden Markov model, the inference model exactly matches the data generation model. In this setting we therefore expect a reasonable good recovery of the



**Figure 2. Empirical transition probabilities versus the HMM transition probabilities.** The figure compares empirical transition probabilities from 25 simulations with the transition probabilities in the hidden Markov model, in the case of 4 states (i.e. divergence times are grouped into four disjoint intervals). The box plot shows the distribution of empirical transition probabilities, with the mean transition probabilities as a blue dot. The red dot shows to the transition probabilities obtained from the CTMC calculations.



**Figure 3. The likelihood of the three parameters.** The plot shows the likelihood as a function of each of the three parameters: divergence time, ancestral effective population time, and recombination rate. For each parameter, the two other parameters are kept fixed at their true value while the one parameter varies. The vertical blue line shows the maximum likelihood estimate while the vertical red line shows the true value.



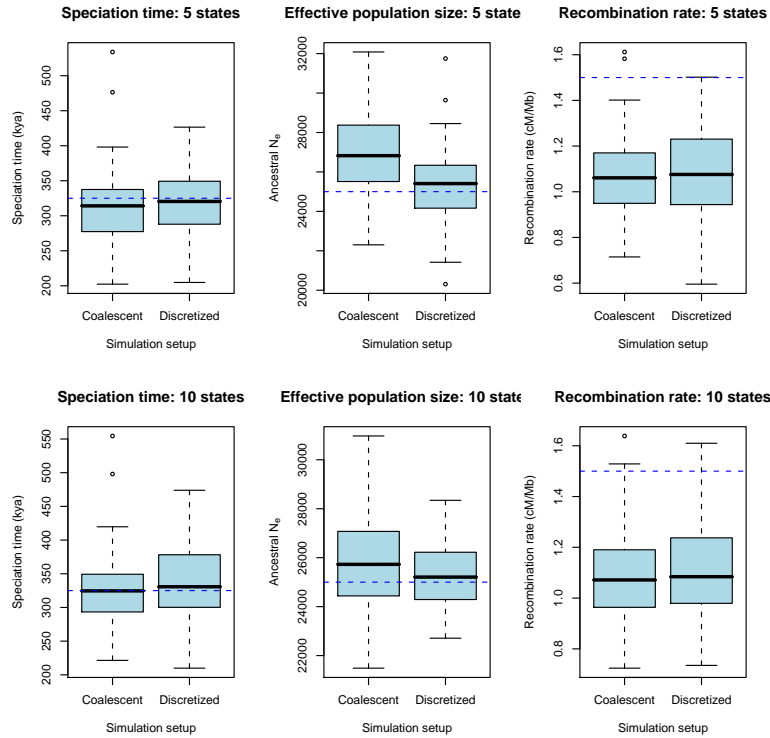
**Figure 4. Parameter estimates when simulating either using the HMM or the coalescent process.** The plot shows the distribution of parameter estimates over 100 simulations from either the coalescent process with recombination or from the HMM. The dashed blue lines show the parameters used for the simulations.

parameters compared to the coalescent process where not all inference assumptions are met.

To compare the inference accuracy between data generated from the coalescent process and from the HMM, we simulated 100 data sets from either process and estimated the parameters for each data set. The results are shown in Fig. 4. As can be seen in the figure, the estimates from data generated with the HMM recovers the parameters accurately, while only the speciation time and the ancestral effective population size is accurately recovered when analyzing data from the coalescent process, while the recombination rate is under-estimated. Thus we must conclude that one or more of the assumptions underlying our approximation to the coalescent process leads us to underestimate the recombination rate, but that our approximation does not significantly affect the speciation time or the effective population size estimates.

#### 1.4.2 Consequences of discrete time intervals

In our model, the emission probabilities for a given interval is determined assuming that the time for the most recent common ancestor is at the mean coalescent point in the interval, calculated from a truncated exponential distribution. In the coalescent process, this is not exactly true. The probability for an alignment column,  $a$ , emitted from interval  $i$  is given by integrating over the coalescent time,  $t$  in



**Figure 5. Parameter estimates when simulating with or without discretized time intervals.** The plot shows the distribution of parameter estimates over 100 simulations from the coalescent process with recombination where the time for the most recent common ancestor is either taken from the coalescent process when generating alignment data, or first fixed to the mean of the corresponding time intervals before generating alignment data. The dashed blue lines show the parameters used for the simulations.

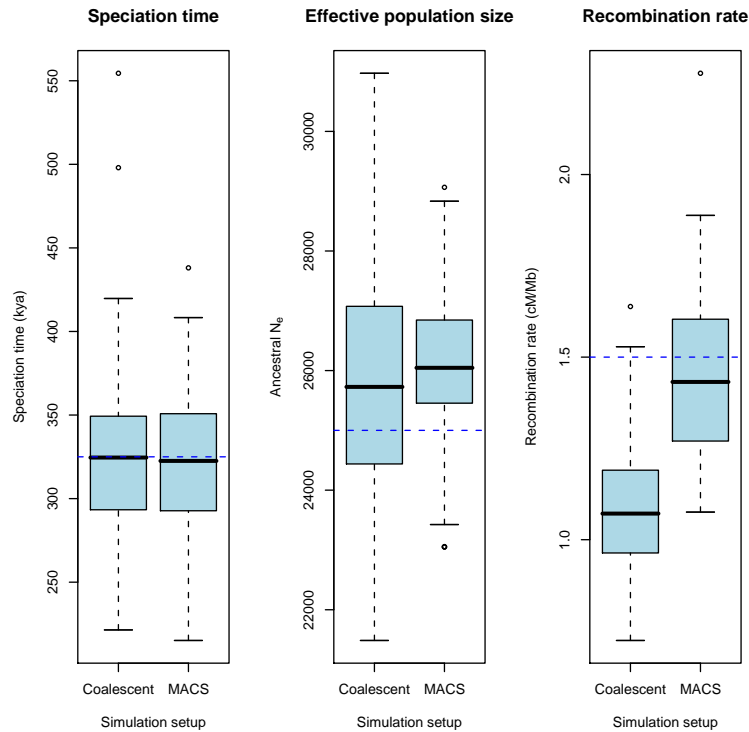
the interval

$$\Pr(a | i) = \int_{\tau_i}^{\tau_{i+1}} \Pr(a | t) p(T = t) dt = E[\Pr(a | T)]$$

where our emission probabilities are given by  $\Pr(a | E[T])$ . Since in general  $E[\Pr(a | T)] \neq \Pr(a | E[T])$  this is a model misspecification.

In our previous work on a CoalHMM based on incomplete lineage sorting (Dutheil *et al.*, Ancestral population genomics: The coalescent hidden Markov model approach, Genetics 2009) we found this to be a source of bias in our estimates of both the divergence time and the ancestral effective population size. To test the consequence of our choice of emission probabilities in the new model, we simulated alignment data by first simulating coalescent times from the coalescent process, then changing each coalescent time to the mean of its corresponding interval, and then generating alignment data from this discretized time coalescent process. Figure 5 shows a comparison of parameter estimates with and without this time-discretization when analyzing the data using five or ten intervals in the HMM.

With five states we see the bias we noticed in our previous work, at least for the effective population size, but for ten states this bias is reduced. The reason for this is that  $E[\Pr(a | T)]$  tends towards  $\Pr(a | E[T])$  as the size of the interval tends to zero.



**Figure 6. Parameter estimates when simulating from the full coalescent process or the Markov coalescence process.** The plot shows the distribution of parameter estimates over 100 simulations from either the coalescent process with recombination or from the coalescence process with the Markov assumption, as implemented in MACS. The dashed blue lines show the parameters used for the simulations.

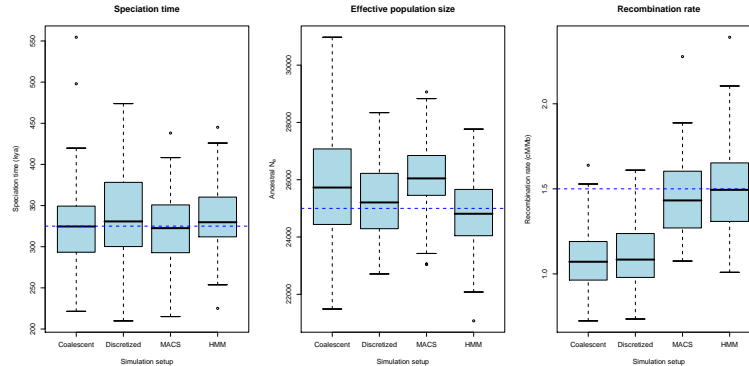
Since for ten states, as we used in the analysis of the orangutan sub-species, the estimates with and without discretized time are very similar, so we conclude that the way we handle time in intervals is not a major source of bias in our estimates.

### 1.4.3 Consequences of the Markov assumption

To test the consequences of the Markov assumption along the alignment, we simulated data from a coalescence process with the Markov assumption (implemented in the simulator MACS; Chen, Marjoram and Wall (2009) Fast and flexible simulation of DNA sequence data, *Genome Res* 19: 136–42). The results are shown in Fig. 6. When the Markov assumption is met, the bias in the estimates of the effective population size is unchanged, but the variance is reduced. The bias in the estimate the recombination rate seems to disappear, however, indicating that this bias is a consequence of the Markov assumption.

### 1.4.4 Substitution rate variability

We simulated 100 ancestral recombination graphs and generated sequence data from these using either a uniform substitution rate or a gamma-distributed rate-across-sites (RAS) model. We then estimated parameters using our HMM with a substitution model using either a uniform or a RAS model for the



**Figure 7. Parameter estimates when simulating with varying models.** The plot shows the distribution of parameter estimates under the different models, varying which of the assumptions are met.

emission probabilities. We saw no significant differences in the estimated parameters between the four combinations of uniform/RAS simulations against uniform/RAS inference.

#### 1.4.5 Conclusions

Drawing conclusions from the model checking, we do not expect the rate variation to influence our estimates, but can identify the assumptions biasing the parameter estimates. Figure 7 shows the four models we have simulated under, and the bias in the three parameters we estimate. For the split time, we do not see any biases, and varying the assumptions of the simulation model does not change this. For the ancestral effective population size, discretizing the time into intervals leads us to over estimate the parameter, and for the recombination rate, assuming the Markov property along the alignment leads us to underestimate the parameter.

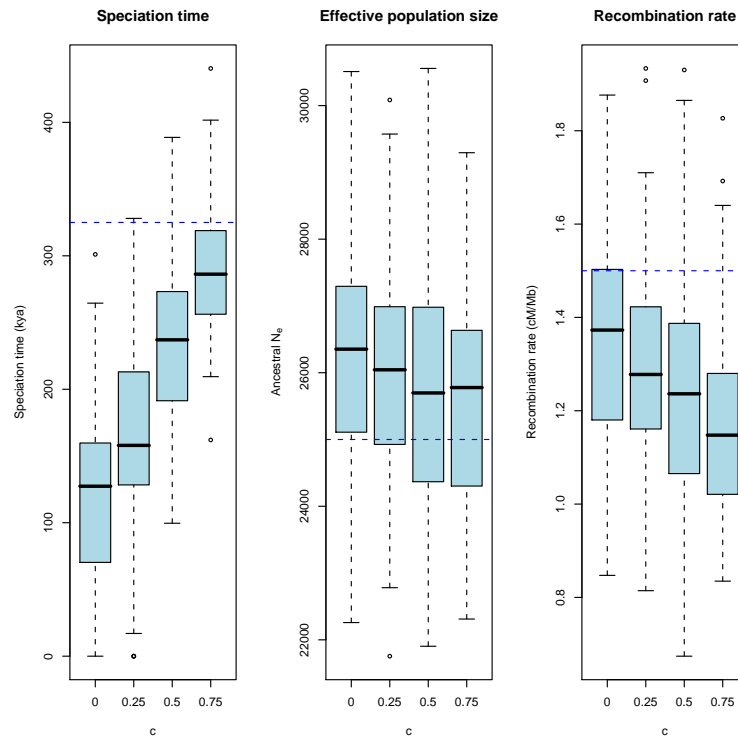
### 1.5 The effect of migration

Our model assumes random mating above the split time and no gene-flow below the split time and thus models a allopatric speciation. To test the effect of this assumption, we simulated data under a more complex speciation model with gene-flow following the split. We simulated data where the ancestral population diverged into two sub-populations at time  $\tau$  exchanged individuals with a (scaled) migration rate  $m$  until time  $\tau \cdot c$  with  $c \in [0, 1]$ , that is the initial population splits at time  $\tau$  but gene-flow does not completely stop until time  $\tau \cdot c$ .

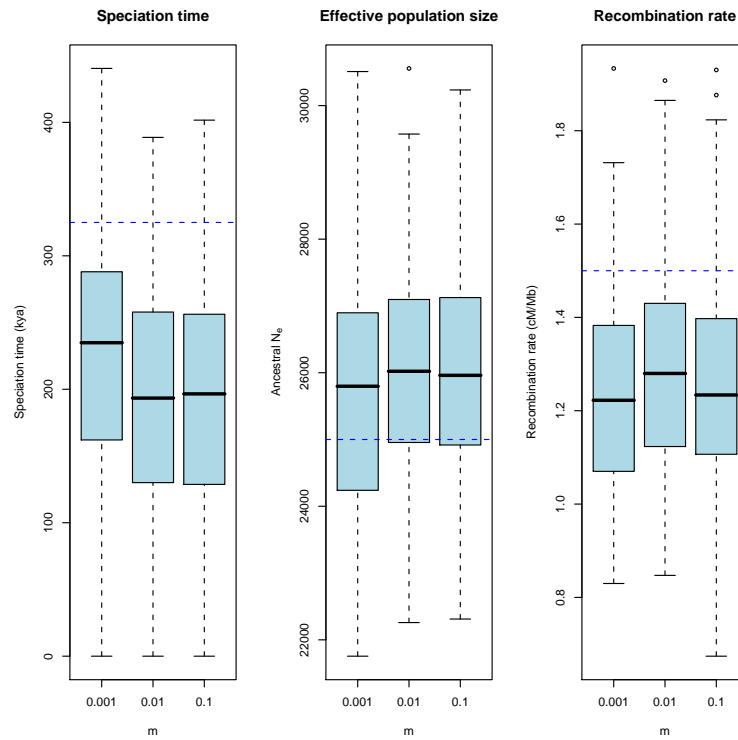
Figure 8 shows the effect on parameter estimates of the time parameter  $c$  while Fig. 9 shows the effect on parameter estimates of the migration parameter  $m$ . The split time is clearly underestimated under this scenario, more so with large migration rates and large time intervals with gene-flow. The effective population size is slightly over-estimated with large migration rates and large time intervals, perhaps not surprising since the coalescence times will have a larger variance in this scenario, and the recombination rate is estimated higher but still biased and under estimating the true rate.

Figure 10 shows the split time estimates plotted against the interval where gene-flow occurs. The estimated split time lies between the two time points where the original population split occurs and where gene-flow stops. For a scenario where there is gene-flow after the original split, our model would therefore likely underestimate the original split but overestimate the end of gene-flow, but for higher migration rates be closer to the latter than the former.

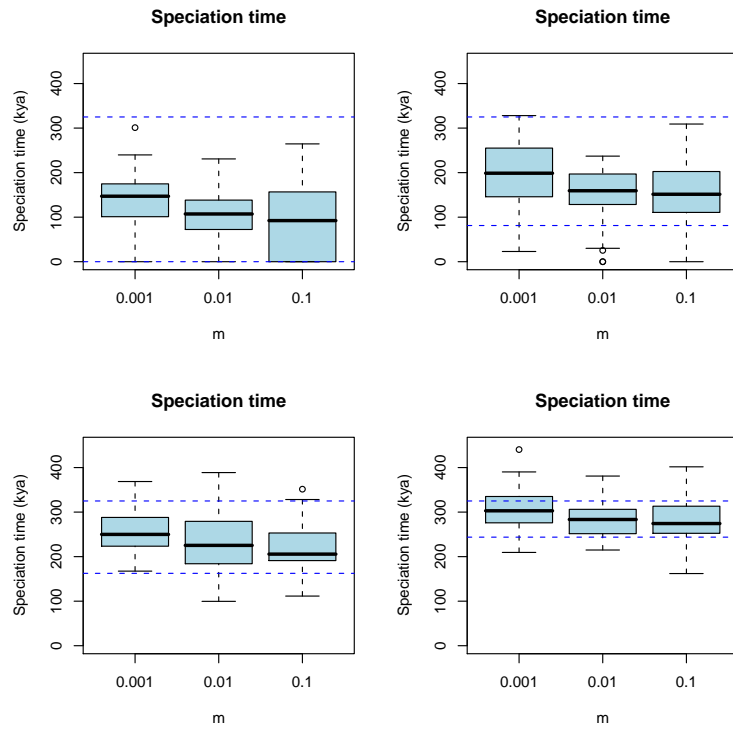




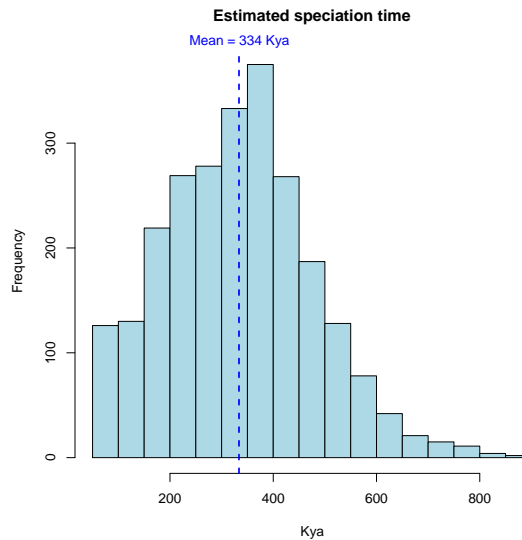
**Figure 8. Parameter estimates versus  $c$  when simulating with gene-flow.** The plot shows the distribution of parameter estimates for varying  $c$ .



**Figure 9. Parameter estimates versus  $m$  when simulating with gene-flow.** The plot shows the distribution of parameter estimates for varying  $m$ .



**Figure 10.** Estimates of the split time as a function of both  $c$  and  $m$ . The plot shows the distribution split time estimates in the model with gene-flow. In each plot, the dashed lines indicate the interval where gene-flow occurs, and the box-plot shows the effect of varying migration rates  $m$ .



**Figure 11. Distribution of sub-speciation time estimates for the entire genome.** The histogram shows the distribution of estimates of the sub-speciation time from the entire genome and the dashed line shows the genome wide mean.

## 2 Analysis of the orangutan sub-species

### 2.1 Distributions of parameter estimates

Figures 11, 12, and 13 shows the distribution of estimated sub-speciation time, ancestral effective population size and recombination rate, respectively.

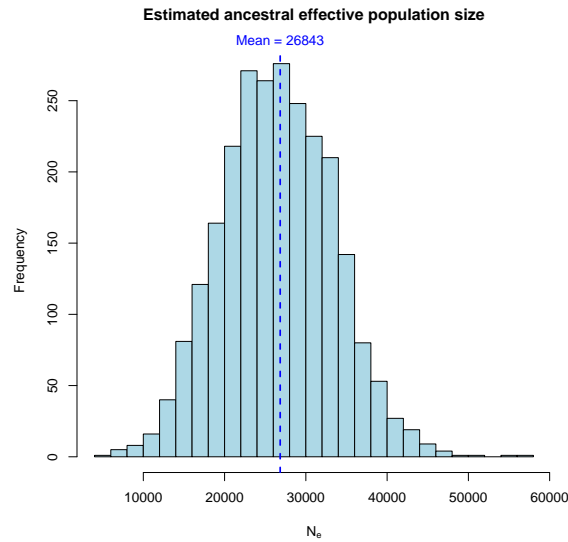
### 2.2 Posterior decoding

Figure 14 shows the posteriors of each of the 10 states for a 100kb fragment of chromosome 22 of the orangutan analysis. The estimated  $N_e$  for this fragment is 29000 and the expected mean coalescent time is thus 1.2 million years. The states in the plot have the most recent coalescent time at the bottom and the most distant coalescent time at the top. The time intervals are chosen to be equi-probable and the break-points between states, in thousands of years since the sub-population split, are 138ky, 294ky, 474ky, 687ky, 947ky, 1283ky, 1757ky and 2567ky.

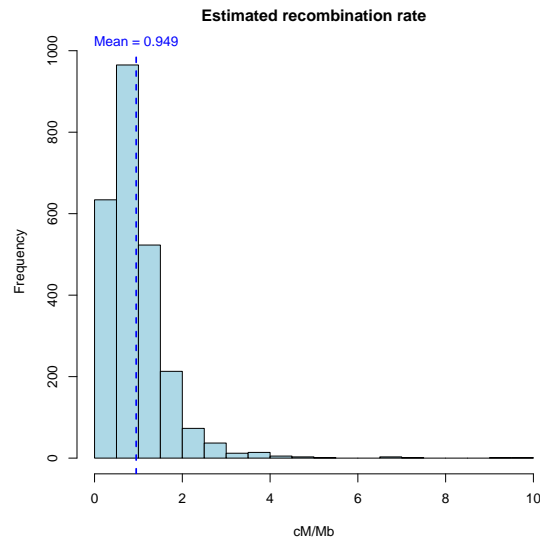
The Figure shows that we are rarely very certain about a specific state but also that the data does contain information about the coalescent times both in the real and in the simulated data.

### 2.3 Robustness of the results

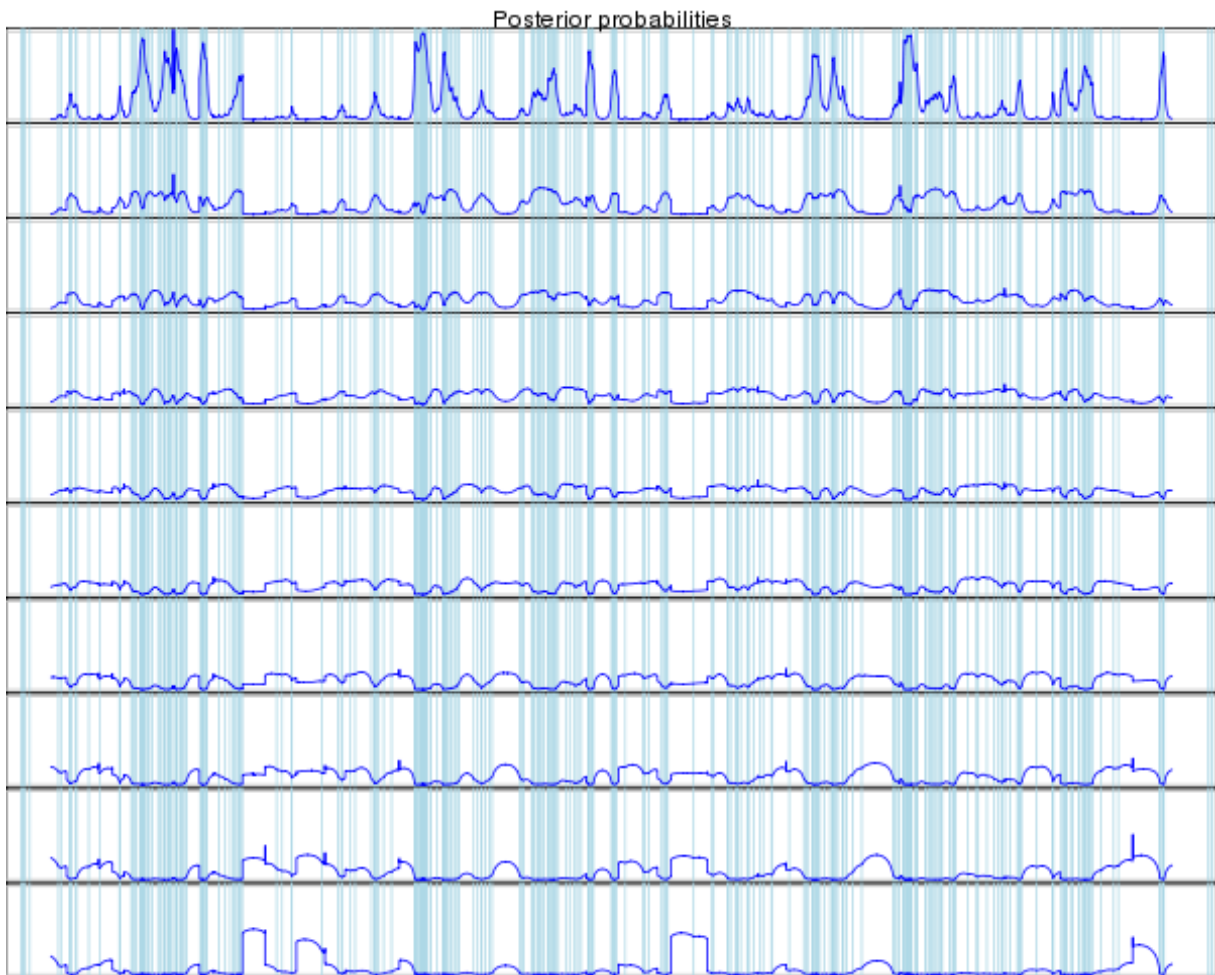
For the two present day effective population sizes, we cannot estimate them independently because of confoundedness, and even setting them to be equal we have very little power to estimate them. This leaves us two options: fixing the present day effective population size and estimating only the ancestral effective population size (the approach we have chosen) or constraining all the three effective population sizes to be equal (and relying on the stronger dependency of the model on the ancestral  $N_e$  to obtain an estimate closer to that than the present day  $N_e$ s).



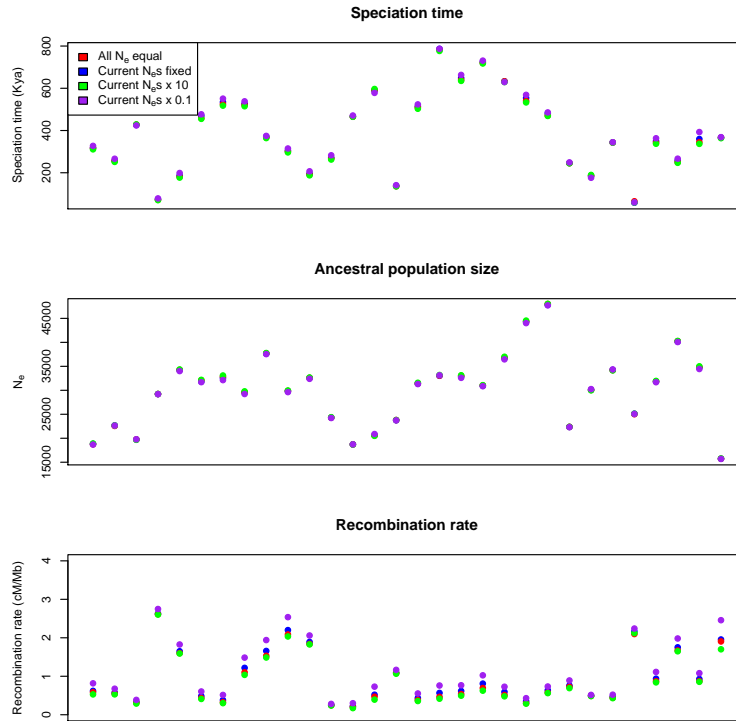
**Figure 12. Distribution of ancestral effective population size for the entire genome.** The histogram shows the distribution of estimates of the ancestral effective population size from the entire genome and the dashed line shows the genome wide mean.



**Figure 13. Distribution of recombination rate for the entire genome.** The histogram shows the distribution of estimates of the recombination rate from the entire genome and the dashed line shows the genome wide mean.



**Figure 14. The posterior decoding of a 10 state analysis.** The plot shows the posterior probability of being in each of the 10 states. The vertical lines show the sites where the nucleotides in the two sub-species differs.



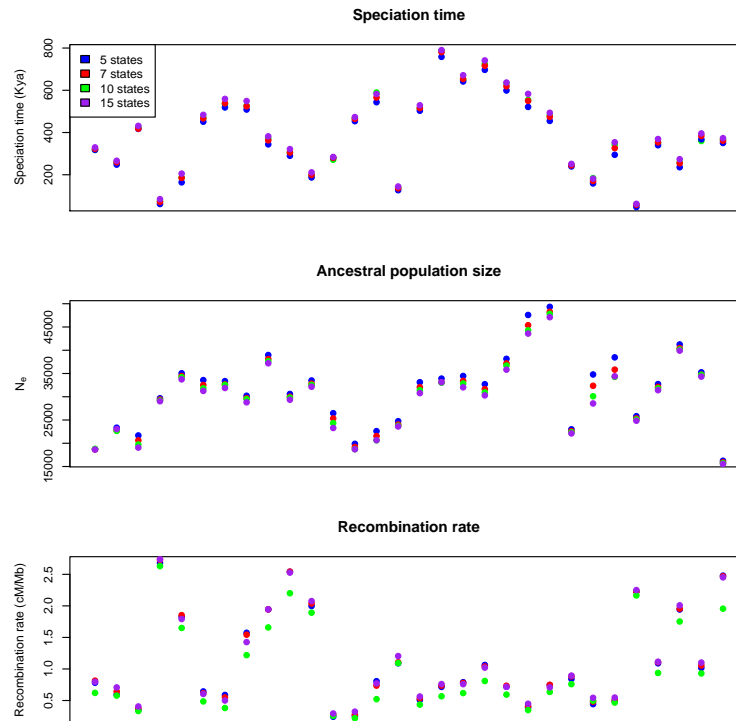
**Figure 15. Robustness to assumptions about the present day effective population sizes.** Estimates for the 30 1Mbp segments from chromosome 22 when varying the assumptions about the present day effective population size.

Testing the robustness to the assumptions about the effective population sizes, we re-analysed chromosome 22 with 1) constraining the  $N_e$ s to be equal, 2) fixing the present day  $N_e$ s to ten times the value we used in the genome wide analysis, and 3) fixing the present day  $N_e$ s to one tenth of the values we used in the genome wide analysis. Figure 15 shows the result. While a paired t-test shows a significant difference between the four setups, clearly the differences in estimates caused by the assumptions about the effective population size is insignificant compared to the variability between segments.

The number of hidden states also affects the estimates, with few states leading to biased estimates of the sub-speciation time and the ancestral effective population size. To test the robustness to the number of states on the real data, we again re-analysed chromosome 22, varying the number of hidden states. Figure 16 shows the result. As with the present day effective population sizes, a paired t-test shows that the number of states changes the estimates, but, again, the changes in estimates caused by changing the number of states is again insignificant to the variability between segments.

Figure 16 shows the model's robustness to the number of hidden states. Again, the changes in estimates caused by varying the number of states is insignificant compared to the variance between 1Mbp segments.

The mean difference in the speciation time, when increasing the number of states, increases with 14,784 when moving from five to seven states, with 2523 when moving from seven to ten states, and with 12,483 when moving from ten to 15 states, consistent with what we observed in Fig. 6 in the main manuscript. In contrast, the standard deviation between different 1Mbp segments for 10 states – the



**Figure 16. Robustness to the number of hidden states.** Estimates for the 30 1Mbp segments from chromosome 22 when the number of hidden states.

results reported in the main paper – is 187,729, an order of magnitude larger than the changes when changing the number of states.

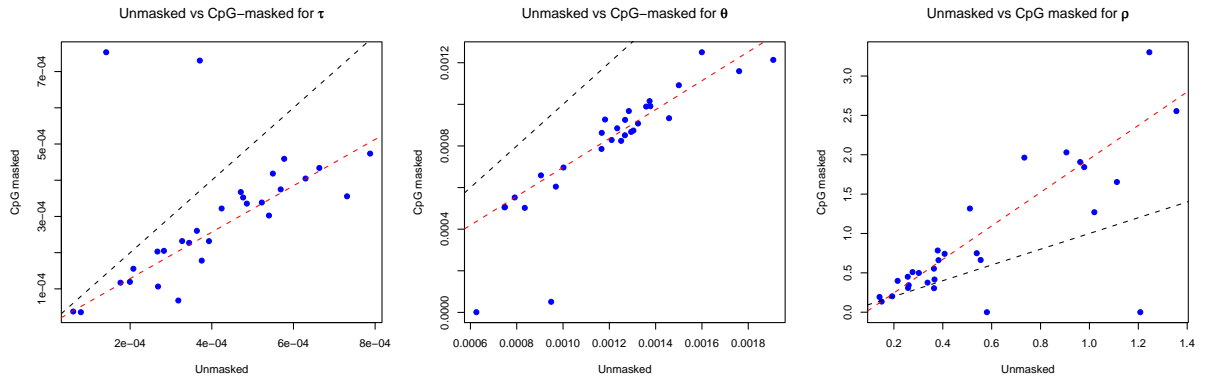
For the effective population size, the mean difference decreases by 898 when moving from five to seven states, decreases by 501 when moving from seven to ten states, and decreases by 429 when moving from ten to 15 states. Again, this is consistent with the simulation results in Fig. 6, and small compared to the variance for the estimates for ten states, where the standard deviation is 7619.

For the recombination rate, the mean difference increases by 0.011 when moving from five to seven states, decreases by 0.15 when moving from seven to ten states, and increases again by 0.16 when moving from ten to 15 states. Based on the simulation results in Fig. 6 we would expect the recombination to decrease when the number of states increase, but for the recombination rate we also observe that the value seems to converge faster towards a fixed mean when the number of states increases, which could explain why we do not see a steady decrease here. Regardless, the difference in estimates when we vary the number of states is small compared to the standard deviation between 1Mbp fragments for ten states, which is 0.68.

## 2.4 Filtering genomic features likely to cause artifacts

Our estimates are based on raw alignments with no filtering of sites. The reasoning being that when analyzing 1Mbp segments, the influence of regions with abnormal evolution – such as sites with a higher mutation rate or sites under selection – will be dwarfed by the influence of sites evolving neutrally and





**Figure 17. Filtering CpG sites.** Correlation between parameter estimates on the segments of chromosome 22 unfiltered or with all CpG sites changed to “unknown nucleotides” (NN). The dashed black line indicates the  $x = y$  line while the dashed red line indicates the best linear fit of the estimates. For the linear fit, two outliers were removed.

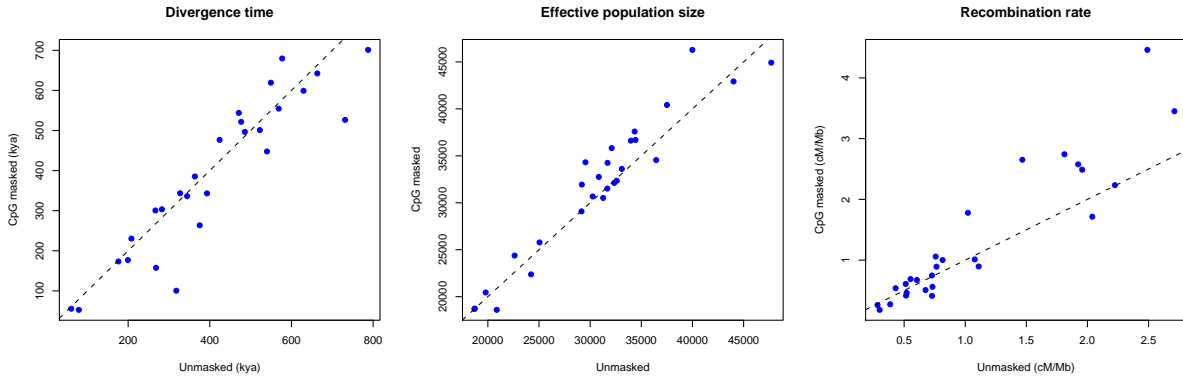
according to our model.

To test this assumption, we filtered out CpG sites — known to have a higher mutation rate — and exons — expected to be under selection — from chromosome 22 and re-ran the analysis.

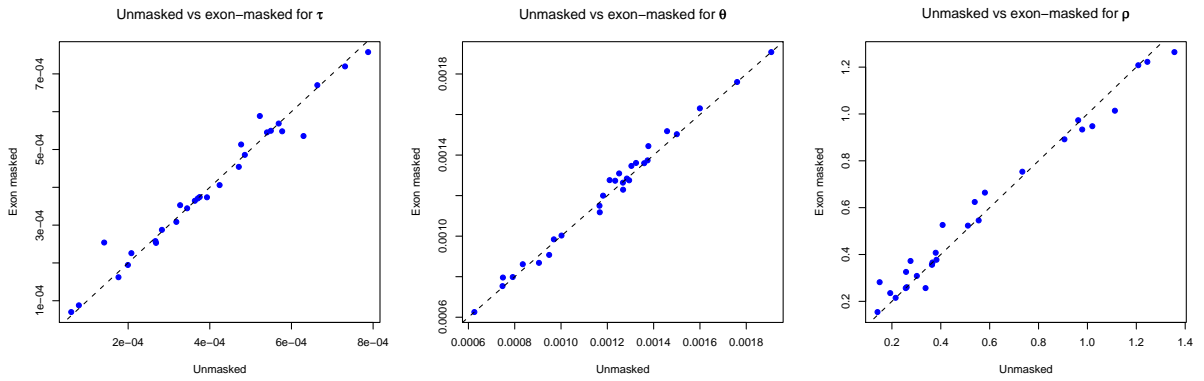
Figure 17 shows the correlation between unmasked results and CpG masked results. With CpG sites masked, the substitution rate of  $10^{-9}$  per site per year will be incorrect since a high fraction of the substitutions will be masked out, so we have not scaled the results as in the paper, but plotted estimates scaled by the expected number of substitutions, and in the plot  $\tau$  denotes the divergence time in number of substitutions,  $\theta$  the scaled effective population size, and  $\rho$  the number of recombinations per substitution.

We see a drop in the divergence and effective population size — as expected since the absolute number of substitutions will be lower when a fraction of substitutions is masked out — but generally a high correlation with the unmasked estimates. Adjusting for the lower substitution rate when accounting for the lack of CpG substitution, we would therefore expect that the scaled estimates are not affected by CpG sites. More than half of the substitutions, 0.68, in the alignments used for the figure are CpG substitutions and rescaling the estimates according to the reduction in substitutions when removing CpG substitutions we get similar results with and without CpG sites masked, see Fig. 18.

The correlation between unmasked estimates and estimates when exons are masked out is shown in Figure 19. Here we see that the potential selection on exons have essentially no effect on the estimates. This matches our expectation that the 1%–2% of sites in exons do not contribute significantly to the estimates in 1Mbp segments.



**Figure 18. Filtering CpG sites and rescaling taking into account the fewer substitutions.** Correlation between parameter estimates on the segments of chromosome 22 unfiltered or with all CpG sites changed to “unknown nucleotides” (NN), when the estimates are scaled with different substitution rates, with  $10^{-9}$  for the unmasked estimates and  $0.68 \cdot 10^{-9}$  for the CpG filtered estimates. The dashed black line indicates the  $x = y$ .



**Figure 19. Filtering exons.** Correlation between parameter estimates on the segments of chromosome 22 unfiltered or with all exon sites changed to “unknown nucleotides” (N). The dashed black line indicates the  $x = y$ .