Supplementary results for: The phylodynamic threshold of measurably evolving populations

Ariane Weber[1,*], Julia Kende[2], Sanni Översti[1,‡] and Sebastian Duchene[3,4,‡,*].

[1] Max Planck Institute of Geoanthropology, Jena, Germany.

[2] Institut Pasteur, Université Paris Cité, Bioinformatics and Biostatistics Hub, Paris, France.

[3] ED-ID unit, Dept of Computational Biology, Institut Pasteur, Paris, France.

[4] Peter Doherty Institute for Infection and Immunity, Dept of Microbiology and Immunology, University of Melbourne, Melbourne, Australia.

*email: weber@gea.mpg.de, sduchene@pasteur.fr

‡ Equal contribution to the supervision of this work.

In fig. S1 we show the simulation results for a situation where the prior on the evolutionary rate is misleading and with most of its density falling on much higher values than those used to generate the data. Here, the prior on the evolutionary rate is $\Gamma(shape = 1.5, rate = 10^4)$ (mean=1.5×10^{-4}, and 95% range from 1.08×10^{-5} to 4.7×10^{-4}), where as that for the population size is an exponential distribution with mean 5,000, and thus concentrated on the true value. Note that the resulting prior on the tree height is not misleading.

In fig. S2 we show the posterior distribution of the tree height for our empirical analyses of Hepatitis B virus with varying sampling window widths, while fig S4 shows the posterior distribution of the tree height for our empirical analyses with temporal sampling bias. The corresponding results with a misleading prior that favours low evolutionary rates and old divergence times are shown in figs **??**, 5. The misleading prior on the evolutionary rate is $\Gamma(shape = 1.5, rate = 10^6)$, while for the population size it is an exponential distribution with mean 50,000.

Finally, figs 6 and 7 show the estimates of the evolutionary rate for these data using the 'misleading' prior.

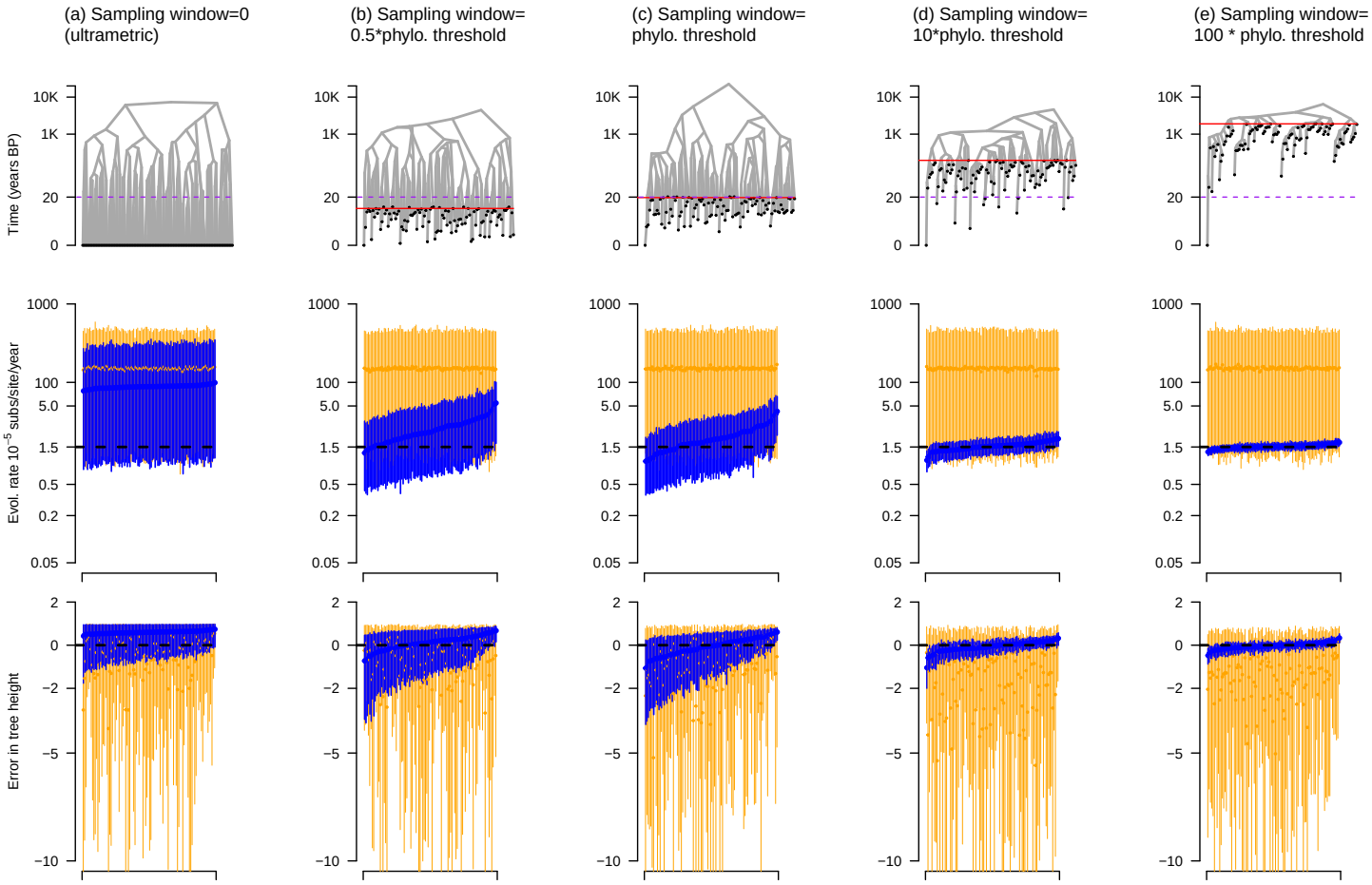

Figure 1: Simulations of varying sampling window widths with a 'misleading prior'. Each column corresponds to a simulation setting: (a) is for ultrametric trees where all samples are collected at the same point in time, (b) is for the situation where the sampling window is 10 years (half the expected phylodynamic threshold), (c) is where the sampling window is exactly the expected phylodynamic threshold of 20 years. Scenarios (d) and (e) denote sampling windows that are 10 and 100 times the expected phylodynamic threshold. The first row is an example of a simulated phylogenetic tree with branch lengths scaled in units of time. The black circles represent genomic samples. The purple dashed line is the expected phylodynamic threshold and the solid red line is for the oldest sample, such that it represents the sampling window. Note that time here is shown in $\log_{10}$ scale. The second row is the estimated evolutionary rate over 100 simulations. The dashed black line is the value used to generate the data (i.e. the ground truth), the dark blue bars are the posterior, and those in orange are the prior. For the prior and the posterior we use solid circles to show the mean estimate and the width of the error bars denotes the 95% quantile range. The third row is the estimate of the error in tree height (the age of the tree). The error in tree height is calculated as $\frac{true-estimated}{true}$.

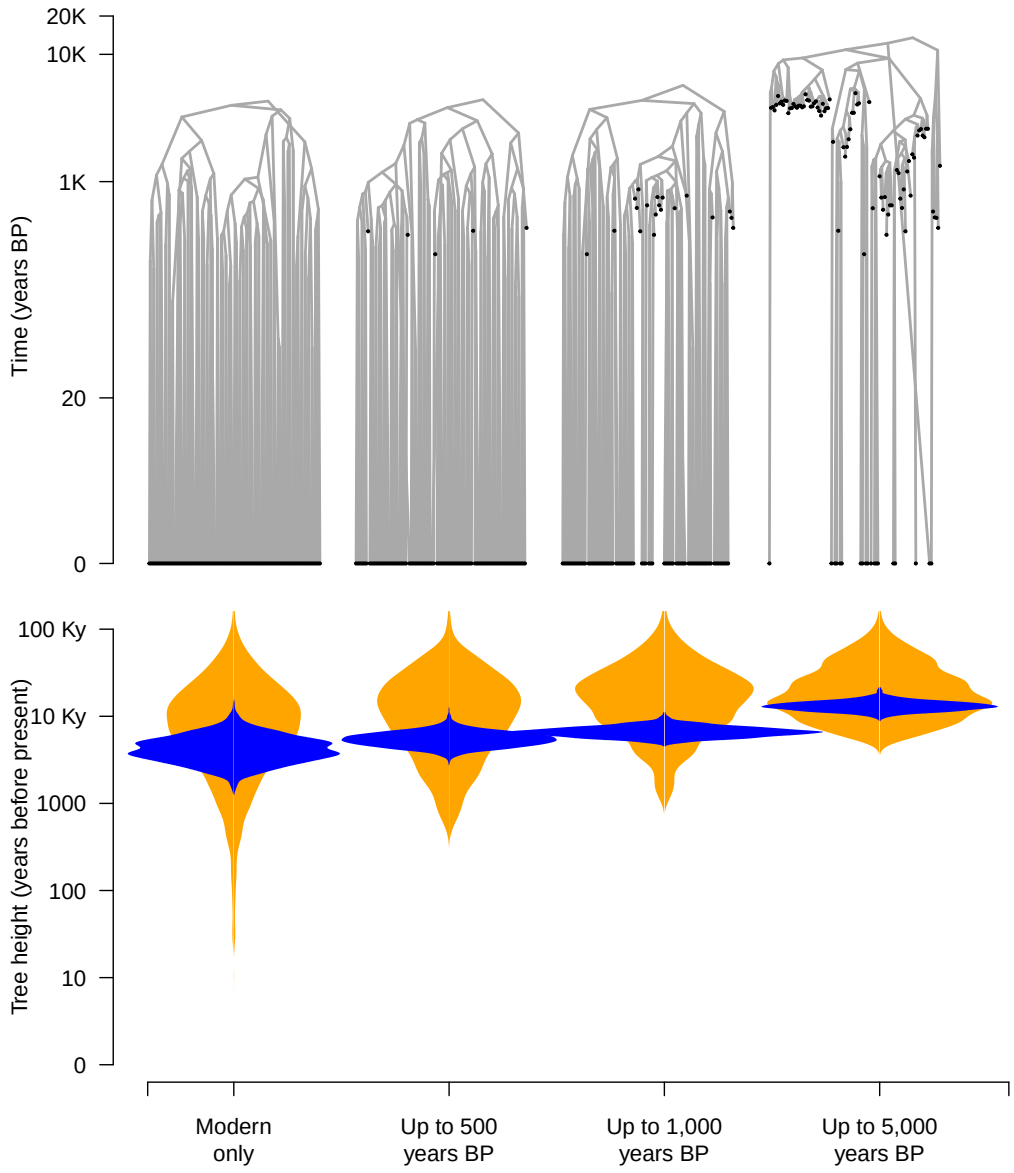

Figure 2: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the tree height in blue and its corresponding prior in orange.

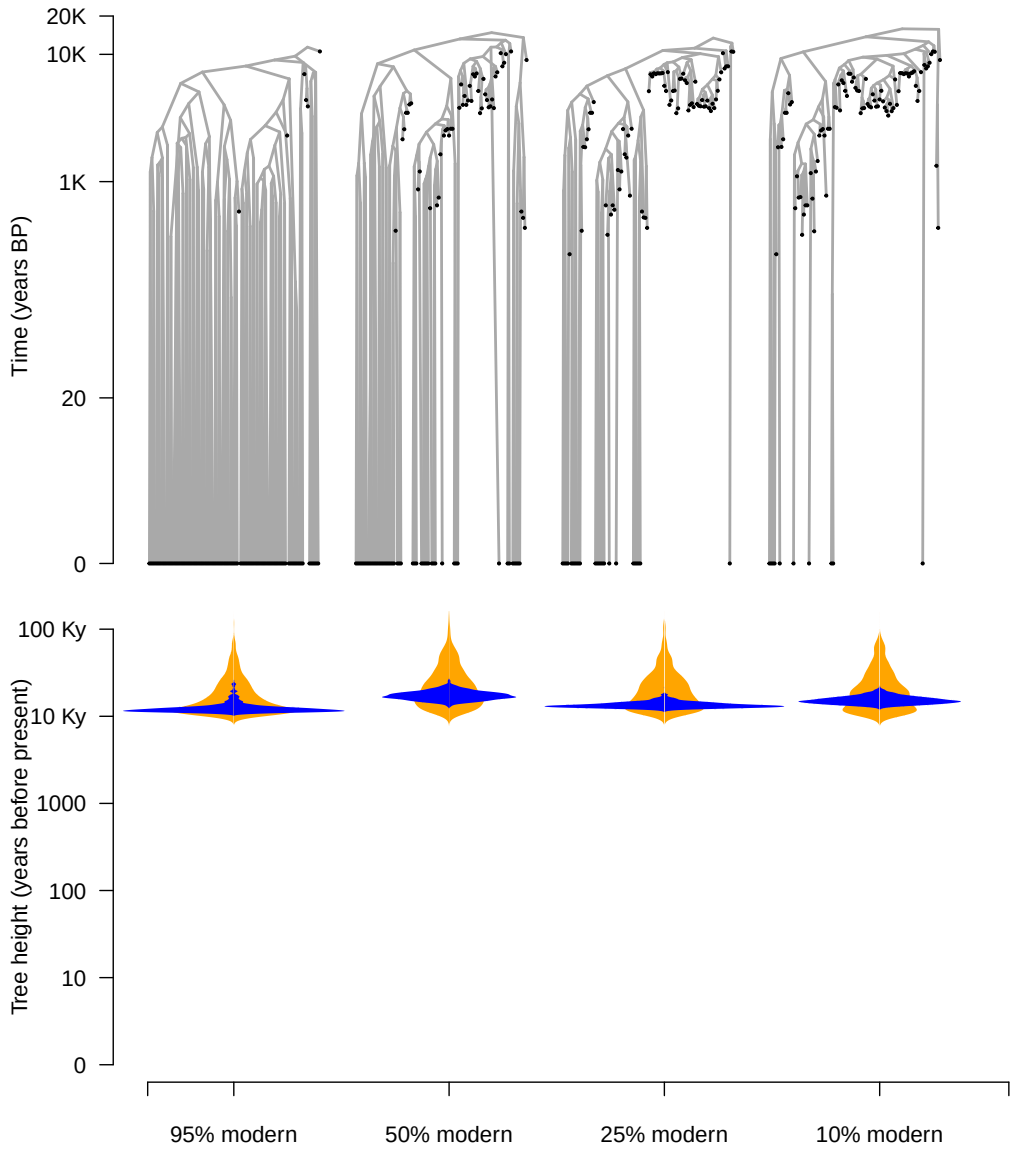

Figure 3: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the tree height in blue and its corresponding prior in orange.

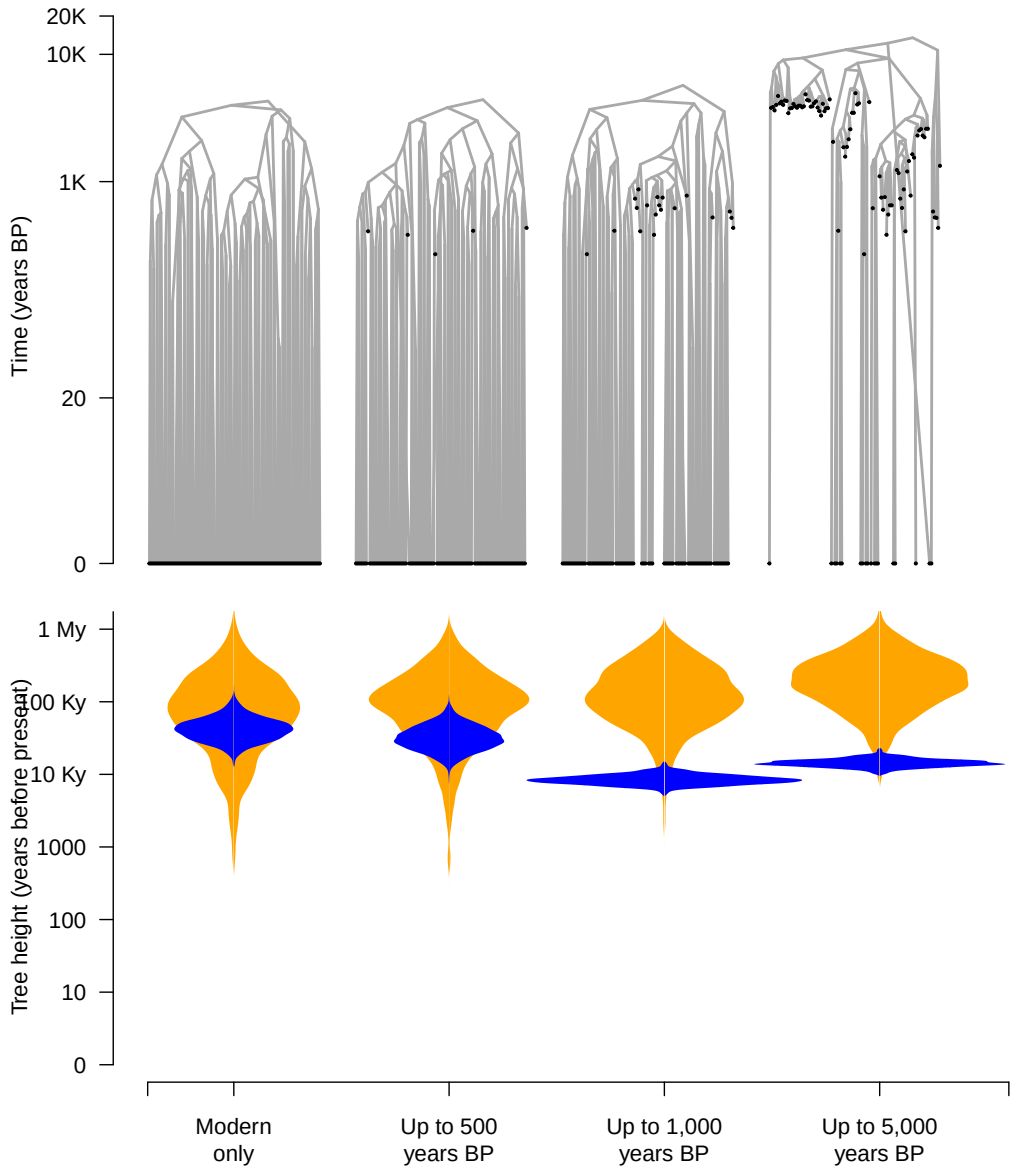

Figure 4: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data using a 'misleading' prior configuration. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the tree height in blue and its corresponding prior in orange.

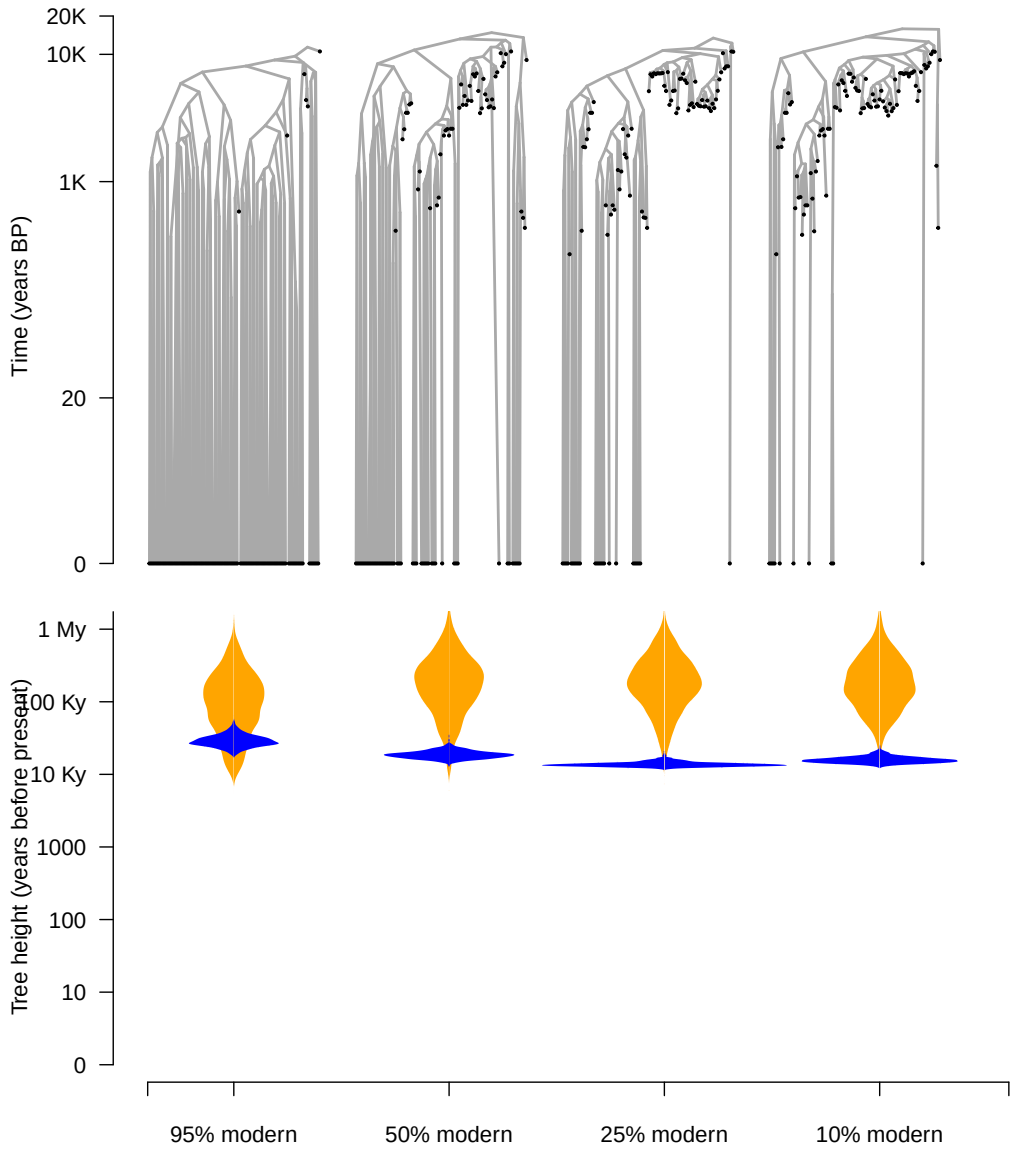

Figure 5: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data using a 'misleading' prior. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the tree height in blue and its corresponding prior in orange.

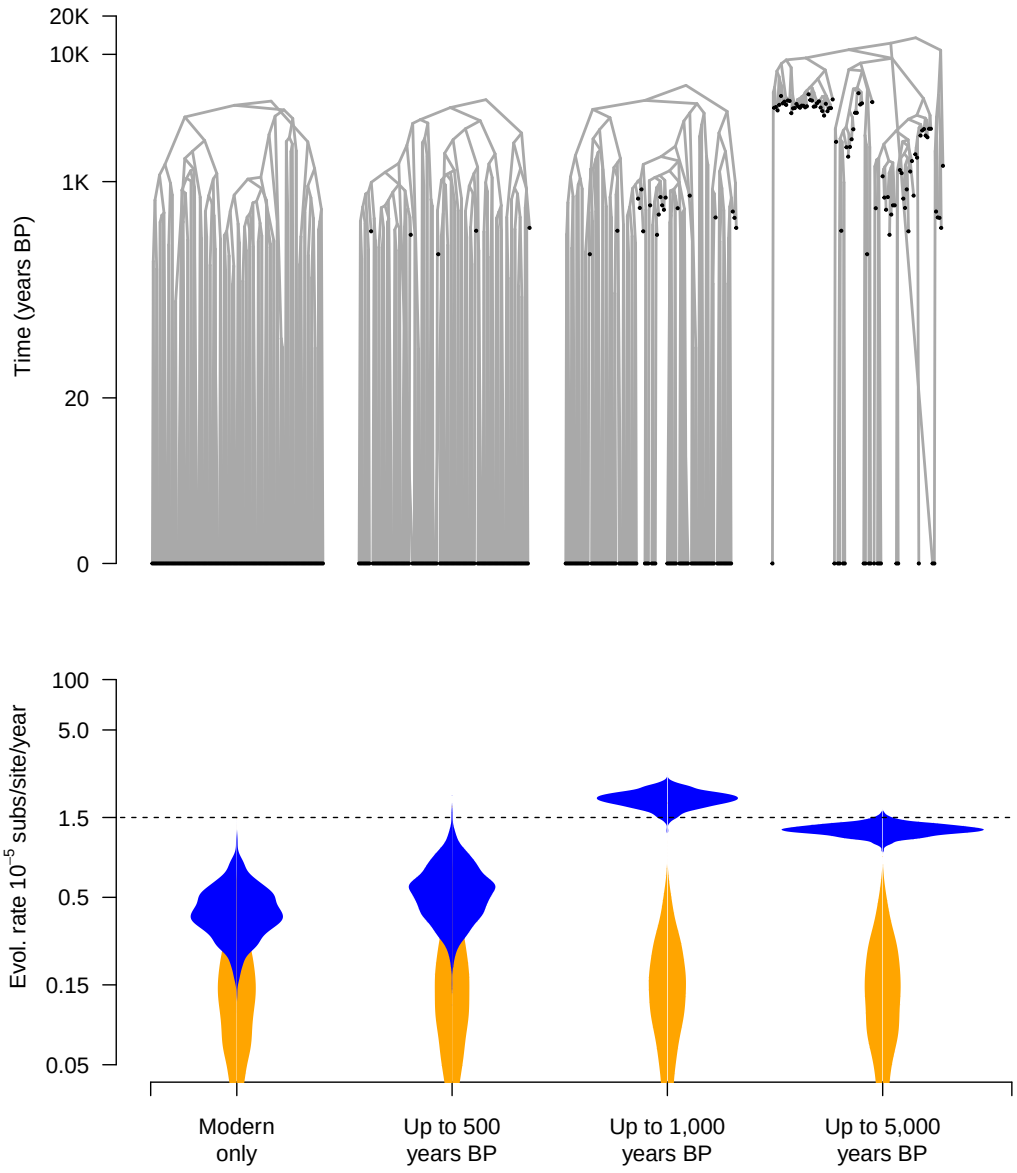

Figure 6: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data under a 'misleading' prior configuration. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the evolutionary rate in blue and its corresponding prior in orange.

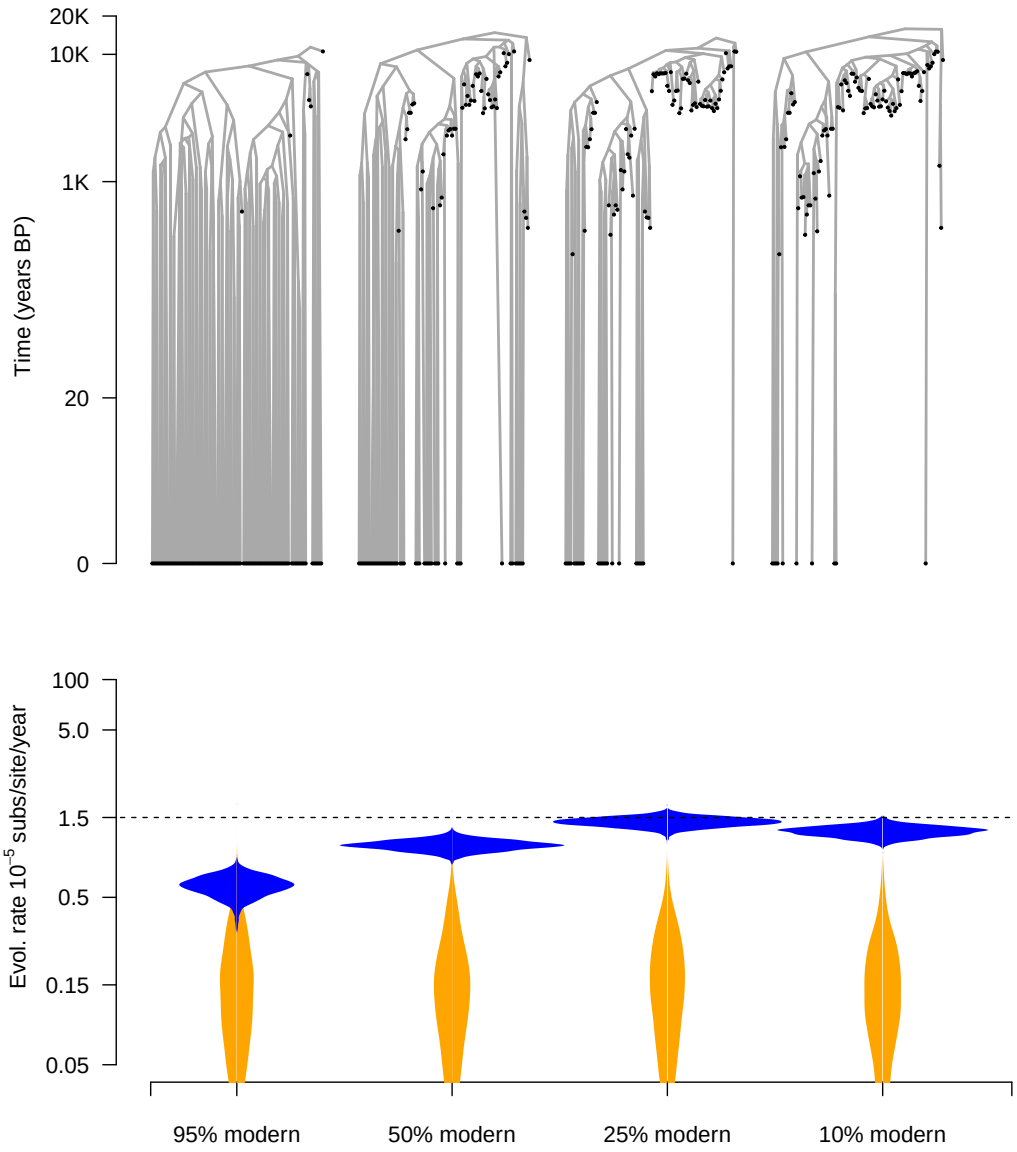

Figure 7: Results from empirical analyses of Hepatitis B virus (HBV) ancient DNA data using a 'misleading' prior. The phylogenetic trees correspond to highest clade credibility trees from three analyses where the data were subsampled to include an increasing number of ancient samples. First, we consider a data set for which the samples are 95% modern and the remaining 5% being the most ancient. Then, we reduce the number of modern samples to 50%, 25% and 10%, and the rest being ancient. Note that the sampling window is constant because we always retain the most ancient samples. In all cases the data sets consist of 100 genome samples. The violin plots show the posterior distribution of the evolutionary rate in blue and its corresponding prior in orange.