Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations

Sebastian Duchene[*,1], Philippe Lemey[2], Tanja Stadler[3], Simon YW Ho[4], David A Duchene[5], Vijaykrishna Dhanasekaran[6], Guy Baele[2]

[1]Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, VIC, Australia.
[2]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Leuven, Belgium.
[3]Department of Biosystems Science and Engineering, ETH Zürich; and Swiss Institute of Bioinformatics; Basel, Switzerland
[4]School of Life and Environmental Sciences, University of Sydney, Sydney, NSW, Australia
[5]Research School of Biology, Australian National University, Canberra, ACT, Australia.
[6]Department of Microbiology, Biomedicine Discovery Institute, Monash University, Melbourne, VIC, Australia.

***Corresponding author:** E-mail: sebastian.duchene@unimelb.edu.au

Abstract (250 words max)

Phylogenetic methods can use the sampling times of molecular sequence data to calibrate the molecular clock, enabling the estimation of substitution rates and time scales for rapidly evolving pathogens and data sets containing ancient DNA samples. A key aspect of such calibrations is whether a sufficient amount of molecular evolution has occurred over the sampling time window, that is, whether the data can be treated as being from a measurably evolving population. Here we investigate the performance of a fully Bayesian evaluation of temporal signal (BETS) in molecular sequence data. The method involves comparing the fit of two models: a model in which the data are accompanied by the actual (heterochronous) sampling times, and a model in which the samples are constrained to be contemporaneous (isochronous). We conduct extensive simulations under a range of conditions to demonstrate that BETS accurately classifies data sets according to whether they contain temporal signal or not, even when there is substantial among-lineage rate variation. We explore the behaviour of this classification in analyses of five data sets: modern samples of *A/H1N1 influenza virus*, the bacterium *Bordetella pertussis*, and coronaviruses from mammalian hosts, and ancient DNA data sets of *Hepatitis B virus* and of mitochondrial genomes of dog species. Our results indicate that BETS is an effective alternative to other measures of temporal signal. In particular, this method has the key advantage of allowing a coherent assessment of the entire model, including the molecular clock and tree prior which are essential aspects of Bayesian phylodynamic analyses.

***Key words*:** Bayesian phylogenetics, ancient DNA, measurably evolving population, marginal likelihood, molecular clock, temporal signal.

Introduction

The molecular clock has become a ubiquitous tool for studying evolutionary processes in rapidly evolving organisms and in data sets that include ancient DNA. In its simplest form, the molecular clock posits that evolutionary change occurs at a predictable rate over time (Zuckerkandl and Pauling 1965). The molecular clock can be calibrated to estimate divergence times by using sampling time information, the timing of known divergence events, or a previous estimate of the substitution rate (Hipsley and Müller 2014). For example, Korber et al. (2000) used sampling times to calibrate the molecular clock and to infer the time of origin of HIV group 1. Their approach consisted of estimating a phylogenetic tree and conducting a regression of the distance from the root to each of the tips as a function of sequence sampling time. In this method, the slope of the regression is an estimate of the substitution rate in substitutions per site per unit of time, the intercept with the time axis is the age of the root node, and the coefficient of determination (R^2) is the degree to which the data exhibit clocklike behaviour (Rambaut et al. 2016). Despite the practicality of root-to-tip regression, its use as a statistical tool for molecular dating has several well-known limitations. In particular, data points are not independent because they have shared ancestry (i.e., internal branches are traversed multiple times) and a strict clocklike behaviour is assumed by necessity.

The past few decades have seen a surge in molecular clock models that explicitly use phylogenetic information. Bayesian methods have gained substantial popularity, largely due to the wide array of complex models that can be implemented and the fact that independent information, including calibrations, can be specified via prior distributions (Nascimento et al. 2017). Of particular importance is the availability of molecular clock models that relax the assumption of strict clock behaviour by explicitly modelling rate variation among lineages (reviewed by Ho and Duchene (2014) and by Bromham et al. (2018)).

Regardless of the methodology used to analyse time-stamped sequence data, a sufficient amount of molecular evolution must have occurred over the sampling time window to allow reliable estimates of substitution rates and timescales. In such cases, the population can be considered to be 'measurably evolving' (Drummond et al. 2003). The degree of 'temporal information' in sequence data is determined by the sequence length, the substitution rate, and the range of available sampling times. Some viruses evolve at a rate of around 5×10^{-3} subs/site/year (Duchene et al. 2014), such that samples collected over a few weeks can be sufficient to calibrate the molecular clock. In more slowly evolving organisms, such as mammals, a sampling window of tens of thousands of years might be necessary; this can be achieved by including ancient DNA sequences (Drummond et al. 2003; Biek et al. 2015).

Testing for temporal signal is an important step for verifying that the molecular clock can be calibrated using the sampling times (Rieux and Balloux 2016). For this purpose a date-randomization test has been proposed that compares actual substitution rate estimates to those obtained by repeatedly permuting the sequence sampling times (Ramsden et al. 2009). A data set is considered to have strong temporal signal if the rate estimated using the correct sampling times does not overlap with those of the permutation replicates (Duchene et al. 2015, 2018; Murray et al. 2015). An implementation of this test is also available that performs the permutation during a single Bayesian inference (Trovão et al. 2015). The interpretation of the date-randomization test is essentially frequentist in nature, which leads to an inconsistent mixture of statistical frameworks when Bayesian phylogenetic methods are used. Moreover, the procedure is not applicable in cases with small numbers of sampling times, owing to the limited number of possible permutations (Duchene et al. 2015).

We propose a full Bayesian model test, which we refer to as BETS (Bayesian Evaluation of Temporal Signal), to assess temporal signal based on previous analyses by Baele et al. (2012). The approach involves quantifying statistical support for two competing models: a model in which the data are accompanied by the actual sampling times (i.e., the data are treated as heterochronous) and a model in which the sampling times are contemporaneous (i.e., the data are treated as isochronous). Therefore, the sampling times are treated as part of the model and the test can be understood as a test of ultrametricity of the phylogenetic tree. If incorporating sampling times improves the statistical fit, then their use for clock calibration is warranted. The crux of BETS, as with Bayesian model selection, is that it requires calculating the marginal likelihood of the model in question. The marginal likelihood measures the evidence for a model given the data, and calculating it requires integration of its likelihood across all parameter values, weighted by the prior (Kass and Raftery 1995).

Because the marginal likelihood is a measure of model evidence, the ratio of the marginal likelihoods of two competing models, known as the Bayes factor, is used to assess support for one model relative to the other. In the case of applying BETS, let M_{het} represent the heterochronous model, M_{iso} the isochronous model, and Y the sequence data, such that $P(Y|M_{\text{het}})$ and $P(Y|M_{\text{iso}})$ are their respective marginal likelihoods. These models differ in the number of parameters; in M_{iso} the substitution rates and times are nonidentifiable, so the rate is fixed to an arbitrary value, while in M_{het} it is a free parameter. Differences in the number of parameters do not need to be taken into account separately, because the marginal likelihood naturally penalizes excessive parameterization. Kass and Raftery (1995) gave guidelines to interpreting Bayes factors, where a (log) Bayes factor $\log(P(Y|M_{\text{het}})) - \log(P(Y|M_{\text{iso}}))$ of at least 5 indicates 'very strong' support for M_{het} over M_{iso}, a value of 3 indicates 'strong' support, and a value of 1 is considered as positive evidence for M_{het} over M_{iso}.

The importance of model selection in Bayesian phylogenetics has prompted the development of various techniques to calculate marginal likelihoods (reviewed by Baele et al. (2014) and by Oaks et al. (2019)). These techniques can be broadly classified into prior-based and/or posterior-based estimators and path-sampling approaches. Prior- and posterior-based estimators, also known as importance sampling, include the widely used harmonic-mean estimator (Newton and Raftery 1994) and the AICM and BICM (Bayesian analogues to the Akaike information criterion and the Bayesian information criterion, respectively) (Raftery et al. 2007). These scores are easy to compute because they only require samples from the posterior distribution as obtained through Markov chain Monte Carlo (MCMC) integration. However, the harmonic-mean estimator has been shown to have unacceptably high variance when the prior is diffuse relative to the posterior, and, together with the AICM, has shown poor performance in practical settings (Baele et al. 2012, 2013). The BICM requires a sample size to be specified for each parameter, which is far from trivial for phylogenetic inference and therefore remains unexplored for such applications.

Path-sampling approaches include path sampling (originally introduced in phylogenetics as 'thermodynamic integration') (Lartillot and Philippe 2006), stepping-stone sampling (Xie et al. 2011), and generalized stepping-stone (GSS) sampling (Fan et al. 2011; Baele et al. 2016). These methods depend on drawing samples using MCMC from a range of power posterior distributions that represent the path from the posterior to the (working) prior, and therefore require additional computation. Another numerical technique that was recently introduced to phylogenetics is nested sampling (NS) (Maturana et al. 2019), which approximates the marginal likelihood by simplifying the marginal-likelihood function from a multi-dimensional to a one-dimensional integral over the cumulative distribution function of the marginal likelihood (Skilling 2006). Fourment et al. (2019) recently compared the accuracy of a range of marginal-likelihood estimation methods and found GSS to be the most accurate, albeit at increased computational cost. Clearly, the reliability of the marginal-likelihood estimator is a key consideration for applying BETS.

We conducted a simulation study to assess the reliability of BETS under a range of conditions that are typical for data sets of rapidly evolving organisms and of those involving ancient DNA. We also analysed five empirical data sets to showcase the performance of the test in practice. Our analyses demonstrate the utility of BETS to provide accurate evaluation of temporal signal across a wide range of conditions.

Results

Simulations of Measurably Evolving Populations

In our simulations we considered sequence data from heterochronous and isochronous trees. Heterochronous trees represent a situation where there is sufficient temporal signal, whereas isochronous trees lack temporal signal altogether. We simulated heterochronous phylogenetic trees under a stochastic birth-death process with between 90 and 110 tips. To generate isochronous trees we used similar settings, but we assumed a single sampling time. We then simulated substitution rates along the trees according to an uncorrelated relaxed clock with an underlying lognormal distribution with a mean of 5×10^{-3} subs/site/unit time and a standard deviation, σ, of 0, 0.1, 0.5, or 1, where σ=0 is equivalent to simulating under a strict clock. We then simulated sequence evolution using an HKY+Γ substitution model, with parameter values similar to those estimated for influenza virus (Hedge et al. 2013), to generate alignments of 10,000 nucleotides.

Our main simulation conditions produced data sets in which about 50% of the sites were variable. We refer to this simulation scenario as (i) 'high substitution rate and wide sampling window', and we considered three other simulation scenarios that involved (ii) a lower substitution rate of 10^{-5} subs/site/unit time, (iii) a narrower sampling window, and (iv) both of the last two conditions. We analysed the sequence data using a strict clock and an uncorrelated relaxed clock with an underlying lognormal distribution (Drummond et al. 2006). We considered three configurations for sampling times: birth-death sampling times, which are correct for the heterochronous data but not for the isochronous data; identical sampling times, which is correct for isochronous data but not for the heterochronous data; and permuted birth-death sampling times, which are incorrect for both heterochronous and isochronous data.

We estimated the log marginal likelihoods of these six combinations of sampling times and clock models using NS and GSS as implemented in BEAST 2.5 (Bouckaert et al. 2019) and BEAST 1.10 (Suchard et al. 2018), respectively. Our BETS approach ranked the models according to their log marginal likelihoods and computed log Bayes factors of the best heterochronous model (M_{het}) compared with the best isochronous model (M_{iso}).

(i) Simulations with High Substitution Rate and Wide Sampling Window

Both NS and GSS correctly classified data as being heterochronous or isochronous in 10 out of 10 simulations, including in the presence of a high degree of among-lineage rate variation (i.e., σ=1; fig. 1 for heterochronous data and supplementary fig. S1, Supplementary Material online, for isochronous data). Although both marginal-likelihood estimators detected temporal signal, NS supported the relaxed clock over the strict clock for three heterochronous data sets simulated without among-lineage rate variation (σ=0) and for six data sets simulated with low among-lineage rate variation (σ=0.1). In the simulations of isochronous data, NS often favoured the relaxed clock over the strict clock when there was low among-lineage rate variation (σ=0.0 and σ=0.1), albeit mostly with log Bayes factors below 5 (supplementary fig. S1, Supplementary Material online). In contrast, GSS always selected the strict clock under these conditions (fig. 1 and supplementary fig. S1, Supplementary Material online).

For the heterochronous data sets, NS and GSS always displayed very strong support for M_{het} over M_{iso}, with log Bayes factors of at least 90. For the isochronous data sets, the log Bayes factors for M_{iso} relative to M_{het} were overall much lower, but still decisive, ranging from 30 to 50. Furthermore, log Bayes factors tended to decline with an increasing degree of among-lineage rate variation in the data. Another important observation is that in the heterochronous data, the relaxed clock was consistently selected over the strict clock when assuming that the data were isochronous, or when the sampling times had been permuted (fig. 1 and supplementary fig. S1, Supplementary Material online). Moreover, the strict clock with permuted sampling times yielded the lowest log marginal likelihoods for heterochronous data. Both of these patterns are likely to be due to an apparently higher degree of among-lineage rate variation when sampling times are misspecified.

(ii) Simulations with Low Substitution Rate and Wide Sampling Window

Our simulations with a low substitution rate of 10^{-5} subs/site/unit time produced data sets that each had about 10 variable sites, which provides very little information for the estimation of evolutionary parameters. Additionally, due to the stochasticity of the simulation process, increased estimator variance between replicates is to be expected given the small number of variable sites. For the heterochronous data sets, GSS selected the heterochronous model with correct dates in at least 7 out of 10 simulation replicates (fig. 2). Across the simulations with different clock models (40 in total), only in five heterochronous data sets did we find models with permuted sampling times to have the highest log marginal likelihoods. For NS, in 11 out of 40 simulations, either isochronous models or those with random sampling times were incorrectly selected when heterochronous data sets were analysed.

Log marginal likelihoods calculated using GSS tended to support models with sampling times (either permuted or those from the birth-death) for the isochronous data, whereas NS appeared to support all models with similar frequencies (supplementary fig. S2, Supplementary Material online). However, a critical feature of the results from the data sets with a low substitution rate is that the log marginal likelihoods for all models were more similar to one another than those for the data sets with high substitution rate (note that the log marginal likelihood scale in fig. 2 is smaller than that in fig. 1). As a case in point, for the isochronous data with σ=0.1 there were log Bayes factors of about 0.1 for the best model with birth-death sampling times relative to those with permuted sampling times. This result indicates that comparing models with permuted sampling times might be useful for determining whether the data are informative about a particular set of sampling times.

(iii) Simulations with High Substitution Rate and Narrow Sampling Window

We conducted a set of simulations similar to those described in scenario (i) but where sequence sampling spanned only the last 10% of the age of the tree (0.5 units of time, compared with 5 units of time for the simulations with a wide sampling window). These conditions reflect those of organisms with deep evolutionary histories and for which samples are available for only a small portion of this time. Since in these trees the samples were collected over a narrower time window, we used a higher sampling probability to obtain about 100 samples, as in our other simulations (see examples of trees in supplementary fig. S3, Supplementary Material online). For these analyses we only considered heterochronous data because the isochronous case is the same as that in scenario (i).

Both GSS and NS showed excellent performance in detecting temporal signal in this scenario, almost always selecting models with correct sampling times. The exceptions to this pattern occurred for one data set with σ=0.5 and for two data sets with σ=1.0 for NS (fig. 3). Differentiating between the strict clock and relaxed clock appeared somewhat more difficult,

249 particularly for NS, where the relaxed clock with correct sampling times yielded log marginal
250 likelihoods very similar to those for the strict clock for data with low among-lineage rate variation
251 (σ of 0.0 or 0.1). Although NS and GSS performed well in these simulations, the log Bayes factors
252 for M_{het} relative to M_{iso} were much lower than those for data with a high substitution rate and a wide
253 sampling window in (i). One obvious example is in the data with σ=0.0, where the mean log Bayes
254 factors for M_{het} over M_{iso} using GSS was 203.15 with a wide sampling window (fig. 1), but only 35.77
255 when sampling spanned a narrow time window (fig. 3).
256
257 *(iv) Simulations with Low Substitution Rate and Narrow Sampling Window*

258 We considered data sets with a narrow sampling window, as in scenario (iii), and with a low
259 substitution rate of 10^{-5} subs/site/unit time, as in scenario (ii). We generated only heterochronous
260 trees under these conditions, because the isochronous case would be the same as that in (ii).
261
262 Estimates of log marginal likelihoods with GSS and NS were very similar among models, with mean
263 log Bayes factors among data sets of less than 1 for the two models with highest marginal
264 likelihoods for GSS (fig. 4). In the data sets with σ=0.0, GSS and NS always preferred a
265 heterochronous model. However, in a few cases (three for GSS and one for NS) the model with
266 permuted sampling times was selected, indicating that temporal signal was not detected. As with
267 the data sets with low substitution rate and constant sampling (ii), the relaxed clock was sometimes
268 preferred over the strict clock, even when the data sets had no rate variation among lineages.
269
270 *Comparison with Root-to-tip Regression*

271 Using a subset of the heterochronous data sets, we conducted root-to-tip regression using
272 phylogenetic trees inferred using maximum likelihood in PhyML 3.1 (Guindon et al. 2010) with the
273 same substitution model as in our BEAST analyses, and with the placement of the root chosen to
274 maximize R^2 in TempEst (Rambaut et al. 2016). We selected data sets generated with a high
275 substitution rate and with both constant and narrow sampling windows. Because GSS and NS
276 correctly detected temporal signal under these conditions, these regressions demonstrate the
277 extent to which this informal regression assessment matches the BETS approach. We did not
278 attempt to provide a thorough benchmarking of the two methods here.
279
280 All regressions had R^2 values that matched our expectation from the degree of among-lineage rate
281 variation, that is, higher values of σ corresponded to lower values of R^2 (fig. 5). The data with a
282 wide sampling window yielded regression slopes ranging from 7.3×10^{-3} to 5.4×10^{-3} subs/site/unit
283 time, which is similar to the substitution rate values used to generate the data. Although the root-
284 to-tip regression is sometimes used to assess temporal signal, it has no cut-off values to confirm
285 temporal signal. This becomes critical when considering the data with a narrow sampling window,
286 for which the R^2 was between 0.13 and 0.02. For example, the regression for a data set with σ=1
287 and narrow sampling window had an R^2 of 0.02, which is sometimes considered sufficiently low as
288 to preclude molecular clock analyses (Rieux and Balloux 2016). However, BETS supported strong
289 temporal structure under a relaxed clock in this data set, with log Bayes factors of 5.48 for this
290 particular data set, which matches the simulation conditions. More importantly, even with such
291 high rate variation, the substitution rate estimated using a relaxed clock and the correct sampling
292 times included the true value used to generate the data (5×10^{-3} subs/site/unit time), with a 95%
293 highest posterior density (HPD) of between 2.15×10^{-3} and 1.90×10^{-2} subs/site/unit time, while the
294 regression slope was 2.22×10^{-2} subs/site/unit time. A key implication of these comparisons is that
295 BETS provides a formal assessment of temporal signal, unlike statistics computed from the
296 regression. Moreover, the root-to-tip regression appears uninformative when the data have been
297 sampled over a narrow time window and there is some rate variation among lineages.
298

Analyses of Empirical Data Sets

We analysed five empirical data sets with similar configurations of sampling times as in our simulation study (Table 1). Two data sets consisted of rapidly evolving pathogens: *A/H1N1 influenza virus* (Hedge et al. 2013) and *Bordetella pertussis* (Bart et al. 2014). We also analysed a data set with highly divergent sequences of coronaviruses (Wertheim et al. 2013), and two data sets with ancient DNA: *Hepatitis B virus* (Patterson Ross et al. 2018), and mitochondrial genomes of dog species (Thalmann et al. 2013). Due to the demonstrated higher accuracy of GSS over NS (Fourment et al. 2019), we applied the BETS approach using the former method only.

The *A/H1N1 influenza virus* data demonstrated clear temporal signal, with the strict clock and relaxed clock with the correct sampling times having the highest log marginal likelihoods, and a log Bayes factor of M_{het} with respect to M_{iso} of 150 (fig. 6). The strict clock had higher support than the relaxed clock for the correct sampling times (log Bayes factor 3.41). Broadly, this result is consistent with previous evidence of strong temporal signal and clocklike behaviour in this data set (Hedge et al. 2013). Using the strict clock with correct sampling times we estimated a substitution rate of 3.37×10^{-3} subs/site/year (HPD: 2.98×10^{-3} to 3.78×10^{-3}).

We detected temporal signal in the *Bordetella pertussis* data set (fig. 6). The relaxed clock with the correct sampling times had the highest log marginal likelihood, with a log Bayes factor relative to the strict clock of 28.86. The log Bayes factor for M_{het} relative to M_{iso} was 47.40. These results echo previous assessments of these data using a date-randomization test (Duchene et al. 2016). We estimated a mean substitution rate using the best model of 1.65×10^{-7} subs/site/year (95% HPD: 1.36×10^{-7} to 2.00×10^{-7}).

Our analyses did not detect temporal signal in the coronavirus data, for which the strict clock and relaxed clock with no sampling times had the highest log marginal likelihoods. The log Bayes factor of M_{het} relative to M_{iso} was -16.82, indicating strong support for the isochronous model. The relaxed clock was supported over the strict clock, with a log Bayes factor of 19.25 (fig. 7). Previous analyses of this data set suggested an ancient origin for this group of viruses, but here the lack of temporal signal precludes any interpretation of our estimates of substitution rates and timescales.

The *Hepatitis B virus* data set included several human genotypes with complete genomes, where 135 were modern sequences collected from 1963 to 2013 and two were ancient samples from human mummies from the 16th century. Previous studies have not found any temporal signal in these data using different approaches, despite the inclusion of ancient sequences. Our estimates of log marginal likelihoods were consistent with a lack of temporal signal, with a log Bayes factor of -101.51 for M_{het} relative to M_{iso}.

The dog mitochondrial genome data contained samples from up to 36,000 years before the present. BETS detected temporal signal in these data, with a log Bayes factor of 38.77 for M_{het} relative to M_{iso}; this result is consistent with that of a date-randomization test in a previous study (Tong et al. 2018). The estimated substitution rate for these data using the best model had a mean of 1.08×10^{-7} subs/site/year (95% HPD: 7.49×10^{-8} to 1.52×10^{-7}).

Discussion

We have proposed BETS, a method that explicitly assesses the statistical support for including sequence sampling times in a Bayesian framework. It is a test of the strength of the temporal signal in a data set, which is an important prerequisite for obtaining reliable inferences in phylodynamic analyses. BETS considers the model ensemble, such that the method can detect temporal signal using models that account for substitution rate variation among lineages. The results of our

analyses demonstrate that our method is effective in a range of conditions, including when the substitution rate is low or when the sampling window represents a small portion of the timespan of the tree.

BETS does not require date permutations, which differentiates it from the widely used date-randomization test for temporal structure. Date-randomization tests address the question of whether a particular association between sequences and sampling times produces estimates different from those obtained from data sets with permuted sampling times (Duchene et al. 2015; Murray et al. 2015). However, such an approach is not a formal test of temporal signal in the data because the permutations do not necessarily constitute an appropriate null model. In contrast, our method does not require permutations and so has the benefit of being robust to using a small number of sampling times.

Accurate calculations of marginal likelihoods are essential for BETS. In our simulation study, we found that GSS and NS correctly assessed the presence and absence of temporal signal in the data under most conditions. The correct clock model was also identified, although in a few instances NS preferred an overparameterized model. Conceivably, using different marginal-likelihood estimators might affect the actual model selected. Murray et al. (2015) also employed a Bayesian model-testing approach using the AICM to assess temporal signal. In their study, the AICM performed well in simulations, but failed to detect temporal signal in empirical data. We attribute this finding to the low accuracy of AICM relative to path-sampling methods (Baele et al. 2012, 2013), and suggest careful consideration of the marginal-likelihood estimator for tests of temporal signal.

A key advantage of BETS is that the complete model is considered, unlike in simpler data-exploration methods such as root-to-tip regression. Specifically, root-to-tip regression is a visual tool for uncovering problems with data quality and to inspect clocklike behaviour, but the absence of appropriate statistics means that there is no clear way of determining whether the data contains temporal information. Consider the regressions in figure 5 for data with a high substitution rate and narrow sampling window. Even when among-lineage rate variation is low (σ =0.1), the data points form a cloud, with a low R^2 of 0.09. However, the apparent 'noise' around the regression line is probably the result of stochasticity in sequence evolution and of the narrow sampling window relative to the age of the root of the tree. In fact, for this particular data set the model with the highest log marginal likelihood is the strict clock with correct sampling times.

In all of our analyses, we ensured that the priors for different models and configurations of sampling times were identical because, as with all Bayesian analyses, model comparison using marginal likelihoods can depend on the choice of prior (Oaks et al. 2019). For example, the tree prior can affect inferences of temporal signal, as it is part of the full model specification. Here we used an exponential-growth coalescent tree prior, which closely matches the demographic dynamics of the birth-death process under which the data were simulated. The effect of using an inappropriate tree prior on tests of temporal signal requires further investigation, but previous studies have suggested that there is only a small impact on estimates of rates and times if the sequence data are informative (Ritchie et al. 2017; Möller et al. 2018).

An interesting finding is that statistical support for isochronous sampling times in truly isochronous data is lower than that for the correct sampling times in truly heterochronous data. This can potentially lead to an increased risk of incorrectly concluding the presence of temporal signal, but we only found this to be a problem in a small number of cases. In particular, in isochronous data simulated with a low substitution rate, and with very few variable sites, the best models were sometimes those that included sampling times, albeit with very low log Bayes factors (e.g.,

supplementary fig. S2, Supplementary Material online). This probably occurs because stochastic error associated with a small amount of evolution leads to low power for model selection.

Permuting sampling times led to poor model fit, as expected. This procedure requires substantial computing requirements, depending on the number of permutations that are performed, and we find that such date permutations are of limited value for model testing when the data are highly informative (e.g., figs. 1 and 3). However, in data sets with very low information content, such as those that were produced by simulation with a low substitution rate here, conducting a small number of date permutations might offer a conservative approach to determining whether model fit and parameter estimates are driven by a particular set of sampling times, as one would expect in the presence of temporal signal.

The nature of the BETS approach means that every parameter in the model has a prior probability, including the substitution rate. Because substitution rates and times are nonidentifiable, it is conceivable that an informative prior on the rate or on the age of an internal node might have a stronger effect than the sampling times on the posterior, for example if the samples span a very short window of time. Such analyses with informative substitution rate priors effectively include several simultaneous sources of calibrating information (i.e., sampling times, internal nodes, and an informative rate prior). Using sampling times in addition to other sources of calibration information might still be warranted if it improves the fit of the model, which can be tested using our proposed method.

Analyses with multiple calibrations can also allow uncertainty in sequence sampling times, especially in data sets that include ancient DNA, where sampling times can be treated as parameters in the model (Shapiro et al. 2011). BETS provides a coherent approach to assess temporal structure in these circumstances, unlike date-randomization tests that typically use point values for sampling times. In fact, BETS can be used as a means to validate whether a sample is modern or ancient.

In general, the uptake of Bayesian model testing in phylogenetics has great potential for improving our confidence in estimates of substitution rates and timescales. The test that we have proposed here, BETS, provides a coherent and intuitive framework to test for temporal information in the data.

Materials and Methods

Simulations

We simulated phylogenetic trees under a stochastic birth-death process using MASTER v6.1 (Vaughan and Drummond 2013), by specifying birth rate λ=1.5, death rate μ=0.5, and sampling rate ψ=0.5. This corresponds to an exponentially growing infectious outbreak with reproductive number R_0=1.5 and a wide sampling window. We set the simulation time to 5 units of time, which corresponds to the time of origin of the process. For isochronous trees, we used similar settings, but instead of using the sampling rate, we sampled each tip with probability ρ=0.5 when the process was stopped after 5 units of time (i.e. μ=1 and ψ=0). Some of our analyses consisted of artificially specifying sampling times for isochronous trees, which we set to those that we would have obtained from a birth-death process with μ=0.5 and ψ=0.5.

In a second set of simulations of heterochronous trees, we generated trees with a narrow sampling window. We specified two intervals for μ and ψ. The first interval spanned 4.5 units of time with μ=1.0 and ψ=0, and the second interval 0.5 units of time with μ=0.1 and ψ=0.9. As a result, the process still had a constant become uninfectious rate ($\mu + \psi$), but samples were only collected in

the second interval. The high sampling rate in the second interval resulted in trees with similar numbers of tips to those with a wide sampling window, but where their ages only spanned 0.5 units of time.

We only considered the simulated trees that contained between 90 and 110 tips. The trees generated in MASTER are chronograms (with branch lengths in units of time), so we simulated substitution rates to generate phylograms (with branch lengths in units of subs/site). To do this we specified the uncorrelated lognormal relaxed clock with a mean rate of 5×10^{-3} or 10^{-5} subs/site/unit time and a standard deviation σ of 0 (corresponding to a strict clock), 0.1, 0.5, or 1. We simulated sequence evolution along these phylograms under the HKY nucleotide substitution model (Hasegawa et al. 1985). We added among-site rate variation using a discretized gamma distribution (Yang 1994, 1996) using Phangorn v2.5 (Schliep 2011) to generate sequence alignments of 10,000 nucleotides. We set the transition-to-transversion ratio of the HKY model to 10 and the shape of the gamma distribution to 1, which is similar to estimates of these parameters in influenza viruses (Duchene et al. 2014; Hedge and Wilson 2014). For each simulation scenario we generated 10 sequence alignments.

Estimation of Marginal Likelihoods Using Nested Sampling

We analysed the data in BEAST 2.5 using the matching substitution model, the exponential-growth coalescent tree prior, the strict clock or relaxed clock, and different configurations of sampling times. We chose the exponential-growth coalescent tree prior, instead of the birth-death tree prior, because it is conditioned on the samples instead of assuming a sampling process; this ensures that the marginal likelihoods for isochronous and heterochronous trees are comparable.

We specified proper priors on all parameters, which is essential for accurate estimation of marginal likelihoods (Baele et al., 2013). In our heterochronous analyses the prior on the substitution rate had a uniform distribution bounded between 0 and 1. We made this arbitrary choice to set a somewhat uninformative prior and because the default prior in BEAST 2.5 is a uniform distribution between 0 and infinity, which is improper. Owing to the nonidentifiability of substitution rates and times, neither can be inferred in the absence of calibrating information, so in our isochronous analyses we fixed the value of the substitution rate to 1. The initial NS chain length was chosen so as to draw 20,000 samples, with 20,000 steps, 32 particles, and a subchain length of 5,000 (note that NS is not equivalent to standard MCMC, nor is the definition of an iteration/step). The chain length and its accompanying sampling frequency were adjusted to obtain effective sample sizes for key parameters of at least 200 (computed in the NS output in BEAST 2.5). Examples of MASTER files and BEAST input files for NS are available online (supplementary data, Supplementary Material online).

Estimation of Marginal Likelihoods Using Generalized Stepping-Stone Sampling

We used BEAST 1.10 with the same model specifications and priors as in BEAST2, except for the prior on the substitution rate, for which we used the approximate continuous-time Markov chain reference prior (Ferreira and Suchard 2008). Because our simulation analyses of GSS and NS differ in this prior, the marginal-likelihood estimates are not directly comparable, so for each simulation we report log Bayes factors of competing models instead of the log marginal likelihoods. The GSS implementation in BEAST 1.10 has two different working priors for the tree generative process: a matching tree prior and a product of exponentials. The latter approach is the most generally applicable and is the one that we used here (Baele et al. 2016).

We used an initial MCMC chain length of 5×10^{7} steps sampling every 5000 steps. After discarding 10% of the samples obtained, the remaining samples were used to construct the working distributions for the GSS analysis. These comprised 100 path steps distributed according to

quantiles from a β distribution with α =0.3, with each of the 101 resulting power posterior inferences running for 5×10^5 iterations. We assessed sufficient sampling for the initial MCMC analysis by verifying that the effective sample sizes for key parameters were at least 200 in Coda v0.19 (Plummer et al. 2006). If this condition was not met, we doubled the length of the MCMC and reduced sampling frequency accordingly. Examples of MASTER files and BEAST input files for GSS are available online (supplementary data, Supplementary Material online).

Analyses of Empirical Data Sets

We downloaded sequence alignments from their original publications (Table 1): complete genomes of *A/H1N1 influenza virus* (Hedge et al. 2013), whole genome sequences of *B. pertussis* (Bart et al. 2014; Duchene et al. 2016), RdRP sequences of coronaviruses (Wertheim et al. 2013), complete genomes of *Hepatitis B virus* (Patterson Ross et al. 2018), and dog mitochondrial genomes (Thalmann et al. 2013). The data and BEAST input files are available in the Supplementary Material online.

Briefly, we used similar settings as in our simulations to estimate marginal likelihoods using GSS. For sequence sampling times we considered the correct sampling times, no sampling times (i.e., isochronous), and permuted sampling times. We also specified tree priors as follows: an exponential-growth coalescent for the *A/H1N2 influenza virus*, *Bordetella pertussis*, coronaviruses, and *Hepatitis B virus* data sets, and a constant-size coalescent for the dog mitochondrial genomes as used by Tong et al. (2018). We again chose the HKY+Γ substitution model, except in the analysis of *Hepatitis B virus* data, for which we used the GTR+Γ model (Tavaré 1986), and in the analysis of the dog data set for which we used the SRD06 substitution model (Shapiro et al. 2006) for coding regions and the GTR+Γ for noncoding regions.

Supplementary Material

Supplementary data are available online.

Funding

SD was supported by an Australian Research Council Discovery Early Career Researcher Award (DE190100805) and an Australian National Health and Medical Research Council grant (APP1157586). PL acknowledges funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 725422-ReservoirDOCS) and the Research Foundation -- Flanders ('Fonds voor Wetenschappelijk Onderzoek -- Vlaanderen', G066215N, G0D5117N and G0B9317N). SYWH was funded by the Australian Research Council (FT160100167). GB acknowledges support from the Interne Fondsen KU Leuven / Internal Funds KU Leuven under grant agreement C14/18/094. VD was supported by contract HHSN272201400006C from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA.

Acknowledgements

Pending.

References

Baele G., Lemey P. 2014. Bayesian model selection in phylogenetics and genealogy-based population genetics. In: Chen M.-H., Kuo L., Lewis P.O., editors. Bayesian Phylogenetics, Methods, Algorithms, and Applications. Boca Raton, Florida: CPC Press. p. 59–93.

Baele G., Lemey P., Bedford T., Rambaut A., Suchard M.A., Alekseyenko A. V. 2012. Improving the accuracy of demographic and molecular clock model comparison while accommodating

549 phylogenetic uncertainty. Mol. Biol. Evol. 29:2157–2167.
550 Baele G., Lemey P., Suchard M.A. 2016. Genealogical working distributions for Bayesian model
551 testing with phylogenetic uncertainty. Syst. Biol. 65:250–264.
552 Baele G., Li W.L.S., Drummond A.J., Suchard M.A., Lemey P. 2013. Accurate model selection of
553 relaxed molecular clocks in bayesian phylogenetics. Mol. Biol. Evol. 30:239–243.
554 Bart M.J., Harris S.R., Advani A., Arakawa Y., Bottero D., Bouchez V., Cassiday P.K., Chiang C.-S.,
555 Dalby T., Fry N.K. 2014. Global population structure and evolution of Bordetella pertussis and
556 their relationship with vaccination. MBio. 5:e01074-14.
557 Biek R., Pybus O.G., Lloyd-Smith J.O., Didelot X. 2015. Measurably evolving pathogens in the
558 genomic era. Trends Ecol. Evol. 30:306–313.
559 Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchene S., Fourment M., Gavryushkina A., Heled J.,
560 Jones G., Kuhnert D., de Maio N. 2019. BEAST 2.5: An Advanced Software Platform for
561 Bayesian Evolutionary Analysis. PLOS Comput. Biol. 15:e1006650.
562 Bromham L., Duchêne S., Hua X., Ritchie A.M., Duchêne D.A., Ho S.Y.W. 2018. Bayesian molecular
563 dating: Opening up the black box. Biol. Rev. 93:1165–1191.
564 Drummond A.J., Ho S.Y.W., Phillips M.J., Rambaut A. 2006. Relaxed phylogenetics and dating with
565 confidence. PLOS Biol. 4:e88.
566 Drummond A.J., Pybus O.G., Rambaut A., Forsberg R., Rodrigo A.G. 2003. Measurably evolving
567 populations. Trends Ecol. Evol. 18:481–488.
568 Duchene S., Duchene D.A., Geoghegan J.L., Dyson Z.A., Hawkey J., Holt K.E. 2018. Inferring
569 demographic parameters in bacterial genomic data using Bayesian and hybrid phylogenetic
570 methods. BMC Evol. Biol. 18:95.
571 Duchene S., Duchene D.A., Holmes E.C., Ho S.Y.W. 2015. The performance of the date-
572 randomization test in phylogenetic analyses of time-structured virus data. Mol. Biol. Evol.
573 32:1895–1906.
574 Duchene S., Holmes E.C., Ho S.Y.W. 2014. Analyses of evolutionary dynamics in viruses are
575 hindered by a time-dependent bias in rate estimates. Proc. R. Soc. London B. 281:20140732.
576 Duchene S., Holt K.E., Weill F.-X., Le Hello S., Hawkey J., Edwards D.J., Fourment M., Holmes E.C.
577 2016. Genome-scale rates of evolutionary change in bacteria. Microb. Genomics. 2:e000094.
578 Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P.O. 2011. Choosing among partition models in Bayesian
579 phylogenetics. Mol. Biol. Evol. 28:523–532.
580 Ferreira M.A.R., Suchard M.A. 2008. Bayesian analysis of elapsed times in continuous-time Markov
581 chains. Can. J. Stat. 36:355–368.
582 Fourment M., Magee A.F., Whidden C., Bilge A., Matsen I. V, Frederick A., Minin V.N. 2019. 19
583 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. Syst.
584 Biol.:syz046.
585 Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms
586 and methods to estimate maximum likelihood phylogenies: assessing the performance of
587 PhyML 3.0. Syst. Biol. 59:307–321.
588 Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of
589 mitochondrial DNA. J. Mol. Evol. 22:160–174.
590 Hedge J., Lycett S.J., Rambaut A. 2013. Real-time characterization of the molecular epidemiology
591 of an influenza pandemic. Biol. Lett. 9:20130331.
592 Hedge J., Wilson D.J. 2014. Bacterial phylogenetic reconstruction from whole genomes is robust to
593 recombination but demographic inference is not. MBio. 5:e02158-14.
594 Hipsley C.A., Müller J. 2014. Beyond fossil calibrations: realities of molecular clock practices in
595 evolutionary biology. Front. Genet. 5:138.
596 Ho S.Y.W., Duchene S. 2014. Molecular-clock methods for estimating evolutionary rates and time
597 scales. Mol. Ecol. 23:5947–5975.
598 Kass R.E., Raftery A.E. 1995. Bayes factors. J. Am. Stat. Assoc. 90:773–795.
599 Korber B., Muldoon M., Theiler J., Gao F., Gupta R., Lapedes A., Hahn B.H., Wolinsky S.,

600 Bhattacharya T. 2000. Timing the ancestor of the HIV-1 pandemic strains. Science. 288:1789–
601 1796.
602 Lartillot N., Philippe H. 2006. Computing Bayes factors using thermodynamic integration. Syst.
603 Biol. 55:195–207.
604 Maturana P., Brewer B.J., Klaere S., Bouckaert R. 2019. Model selection and parameter inference in
605 phylogenetics using Nested Sampling. Syst. Biol. 68:219–233.
606 Möller S., du Plessis L., Stadler T. 2018. Impact of the tree prior on estimating clock rates during
607 epidemic outbreaks. Proc. Natl. Acad. Sci. USA. 115:4200–4205.
608 Murray G.G.R., Wang F., Harrison E.M., Paterson G.K., Mather A.E., Harris S.R., Holmes M.A.,
609 Rambaut A., Welch J.J. 2015. The effect of genetic structure on molecular dating and tests for
610 temporal signal. Methods Ecol. Evol. 7:80–89.
611 Nascimento F.F., dos Reis M., Yang Z. 2017. A biologist's guide to Bayesian phylogenetic analysis.
612 Nat. Ecol. Evol. 1:1446.
613 Newton M.A., Raftery A.E. 1994. Approximate Bayesian inference with the weighted likelihood
614 bootstrap. J. R. Stat. Soc. Ser. B. 56:3–26.
615 Oaks J.R., Cobb K.A., Minin V.N., Leaché A.D. 2019. Marginal likelihoods in phylogenetics: a review
616 of methods and applications. Syst. Biol. 68:681–697.
617 Patterson Ross Z., Klunk J., Fornaciari G., Giuffra V., Duchene S., Duggan A.T., Poinar D., Douglas
618 M.W., Eden J.-S., Holmes E.C., Poinar H.N. 2018. The paradox of HBV evolution as revealed
619 from a 16thcentury mummy. PLoS Pathog. 14:e1006887.
620 Plummer M., Best N., Cowles K., Vines K. 2006. CODA: Convergence diagnosis and output analysis
621 for MCMC. R News. 6:7–11.
622 Raftery A., Newton M., Satagopan J., Krivitsky P. 2007. Estimating the integrated likelihood via
623 posterior simulation using the harmonic mean identity. Bayesian Statistics. Oxford University
624 Press. p. 1–45.
625 Rambaut A., Lam T.T., Carvalho L.M., Pybus O.G. 2016. Exploring the temporal structure of
626 heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2:vew007.
627 Ramsden C., Holmes E.C.C., Charleston M.A.A. 2009. Hantavirus evolution in relation to its rodent
628 and insectivore hosts: no evidence for codivergence. Mol. Biol. Evol. 26:143–153.
629 Rieux A., Balloux F. 2016. Inferences from tip-calibrated phylogenies: a review and a practical guide.
630 Mol. Ecol. 25:1911–1924.
631 Ritchie A.M., Lo N., Ho S.Y.W. 2017. The impact of the tree prior on molecular dating of data sets
632 containing a mixture of inter-and intraspecies sampling. Syst. Biol. 66:413–425.
633 Schliep K.P. 2011. phangorn: Phylogenetic analysis in R. Bioinformatics. 27:592–593.
634 Shapiro B., Ho S.Y.W., Drummond A.J., Suchard M.A., Pybus O.G., Rambaut A. 2011. A Bayesian
635 phylogenetic method to estimate unknown sequence ages. Mol. Biol. Evol. 28:879–887.
636 Shapiro B., Rambaut A., Drummond A.J. 2006. Choosing appropriate substitution models for the
637 phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7–9.
638 Skilling J. 2006. Nested sampling for general Bayesian computation. Bayesian Anal. 1:833–859.
639 Suchard M.A., Lemey P., Baele G., Ayres D.L., Drummond A.J., Rambaut A. 2018. Bayesian
640 phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 4:vey016.
641 Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect.
642 Math. Life Sci. 17:57–86.
643 Thalmann O., Shapiro B., Cui P., Schuenemann V.J., Sawyer S.K., Greenfield D.L., Germonpré M.B.,
644 Sablin M. V, López-Giráldez F., Domingo-Roura X. 2013. Complete mitochondrial genomes of
645 ancient canids suggest a European origin of domestic dogs. Science. 342:871–874.
646 Tong K.J., Duchêne D.A., Duchêne S., Geoghegan J.L., Ho S.Y.W. 2018. A comparison of methods
647 for estimating substitution rates from ancient DNA sequence data. BMC Evol. Biol. 18:70.
648 Trovão N.S., Baele G., Vrancken B., Bielejec F., Suchard M.A., Fargette D., Lemey P. 2015. Host
649 ecology determines the dispersal patterns of a plant virus. Virus Evol. 1:vev016.
650 Vaughan T.G., Drummond A.J. 2013. A stochastic simulator of birth–death master equations with

application to phylodynamics. Mol. Biol. Evol. 30:1480–1493.
Wertheim J.O., Chu D.K.W., Peiris J.S.M., Pond S.L.K., Poon L.L.M. 2013. A case for the ancient origin of coronaviruses. J. Virol. 87:7039–7045.
Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.
Yang Z. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.
Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11:367–372.
Zuckerkandl E., Pauling L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson V., Vogel H., editors. Evolving Genes and Proteins. New York: Academic press. p. 97–166.

662 **Figure Legends**

663 **FIG. 1.** Log Bayes factors of heterochronous data simulated with a high substitution rate and wide
664 sampling window. Each panel shows the results for data sets simulated with a different degree of
665 among-lineage rate variation, governed by the standard deviation σ of a lognormal distribution. In
666 each plot the *x*-axis depicts six analysis settings, with two clock models, strict clock (SC) and the
667 uncorrelated relaxed clock with an underlying lognormal distribution (UCLN), and three settings for
668 sampling times: generated under the birth-death process (BD), identical sampling times
669 (Isochronous), and permuted (Permuted). The points have been jittered along the *x*-axis to facilitate
670 visualization. The *y*-axis shows log Bayes factors relative to the best model. Red points correspond
671 to estimates using generalized stepping-stone sampling and blue points correspond to estimates
672 using nested sampling. We conducted 10 simulation replicates, with each replicate data set
673 analysed under the six analysis settings and two marginal-likelihood estimators, such that
674 stochastic error might cause differences in the preferred model. The number next to each cloud of
675 points denotes the number of times (out of 10) that the corresponding model had the highest
676 marginal likelihood with generalized stepping-stone sampling (red) and nested sampling (blue).
677
678 **FIG. 2.** Log Bayes factors of heterochronous data simulated under a low substitution rate and a wide
679 sampling window. Symbols and colours are the same as those in figure 1.
680
681 **FIG. 3.** Log Bayes factors of heterochronous data simulated under a high substitution rate and
682 narrow sampling window. Symbols and colours are the same as those in figure 1.
683
684 **FIG. 4.** Log Bayes factors of heterochronous data simulated under a low substitution rate and
685 narrow sampling window. Symbols and colours are the same as those in figure 1.
686
687 **FIG. 5.** Root-to-tip regressions for a subset of data sets simulated with varying degrees of among-
688 lineage rate variation (governed by the standard deviation σ of a lognormal distribution), using a
689 high substitution rate and either a wide or narrow sampling window. The *y*-axis is the root-to-tip
690 distance and the *x*-axis is the time from the youngest tip, where 0 is the present. Each point
691 corresponds to a tip in the tree and the solid line is the best-fit linear regression using least-squares.
692 The coefficient of determination, R^2, is shown in each case. For comparison, the log Bayes factors of
693 the best heterochronous model relative the best isochronous model, BF(M_{het}-M_{iso}), are also shown.
694
695 **FIG. 6.** Log marginal likelihoods estimated using generalized stepping-stone sampling for six
696 analysis settings for sequence data from rapidly evolving pathogens, *A/H1N1 Human influenza virus*
697 and *Bordetella pertussis*. The *y*-axis is the marginal likelihood and the *x*-axis shows the analysis
698 settings, with two clock models, strict clock (SC) and the uncorrelated relaxed clock with an
699 underlying lognormal distribution (UCLN), and three settings for sampling times: generated under
700 the birth-death process (BD), identical sampling times (Isochronous), and permuted (Permuted).
701 Solid points and dashed lines correspond to the log marginal-likelihood estimates. The asterisk
702 denotes the model with the highest marginal likelihood.
703
704 **FIG. 7**. Log marginal likelihoods estimated using generalized stepping-stone sampling for six
705 analysis settings for data sets with ancient DNA or highly divergent sequences. The *y*-axis is the
706 marginal likelihood and the *x*-axis shows the analysis settings, with two clock models, strict clock
707 (SC) and the uncorrelated relaxed clock with an underlying lognormal distribution (UCLN), and
708 three settings for sampling times: generated under the birth-death process (BD), identical sampling
709 times (Isochronous), and randomized (Random). Solid points and dashed lines correspond to the log
710 marginal-likelihood estimates. The asterisk denotes the model with the highest marginal likelihood.
711
712

Tables

Table 1. Details of empirical data sets used in this study.

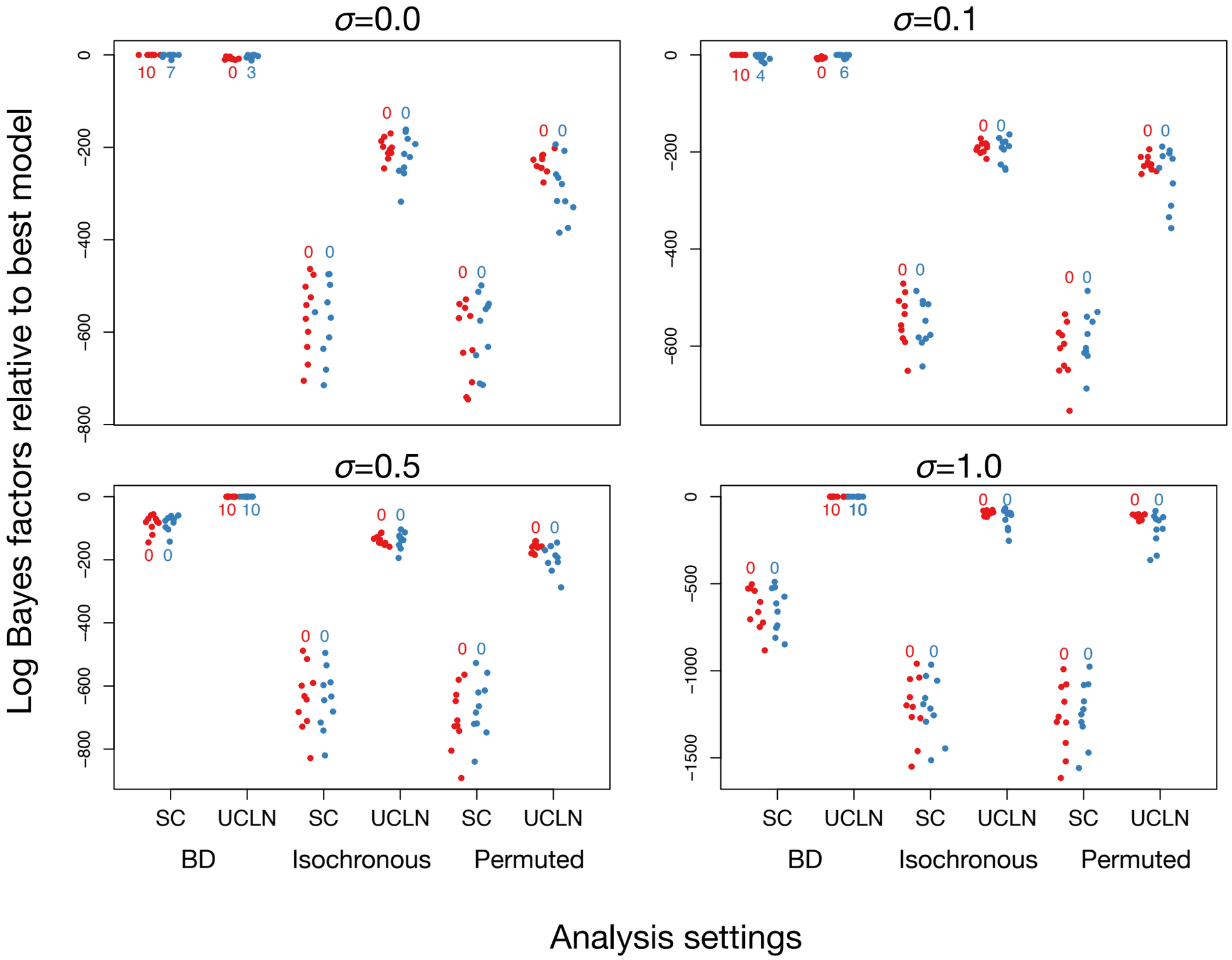
Data set	Number of sites (nucleotides)	Number of samples	Sampling time range	Reference
A/H1N1 influenza virus	13,154	329	10 months (March to December 2009)	Hedge et al. (2013)
Bordetella pertussis	4.9×10^{6}	150	89 years (1920 to 2009)	Bart et al. (2014)
Coronaviruses	1,860	43	70 years (1941 to 2011)	Wertheim et al. (2013)
Hepatitis B virus	3,271	137	445 years (2103 to 1568)	Patterson Ross et al. (2018)
Dog mtDNA	14,596	50	36,000 years (to the present)	Thalmann et al. (2013)

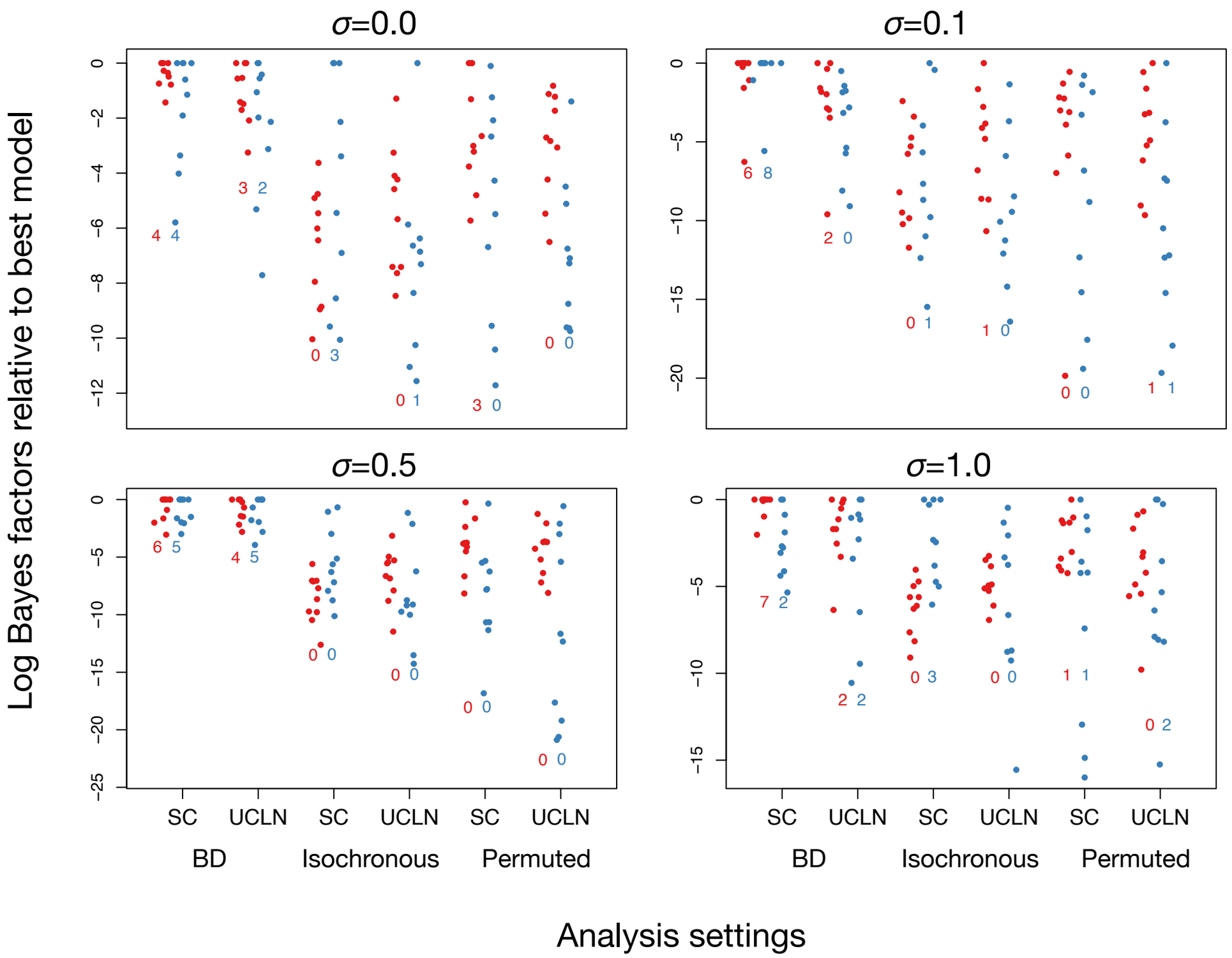
Supplementary Material

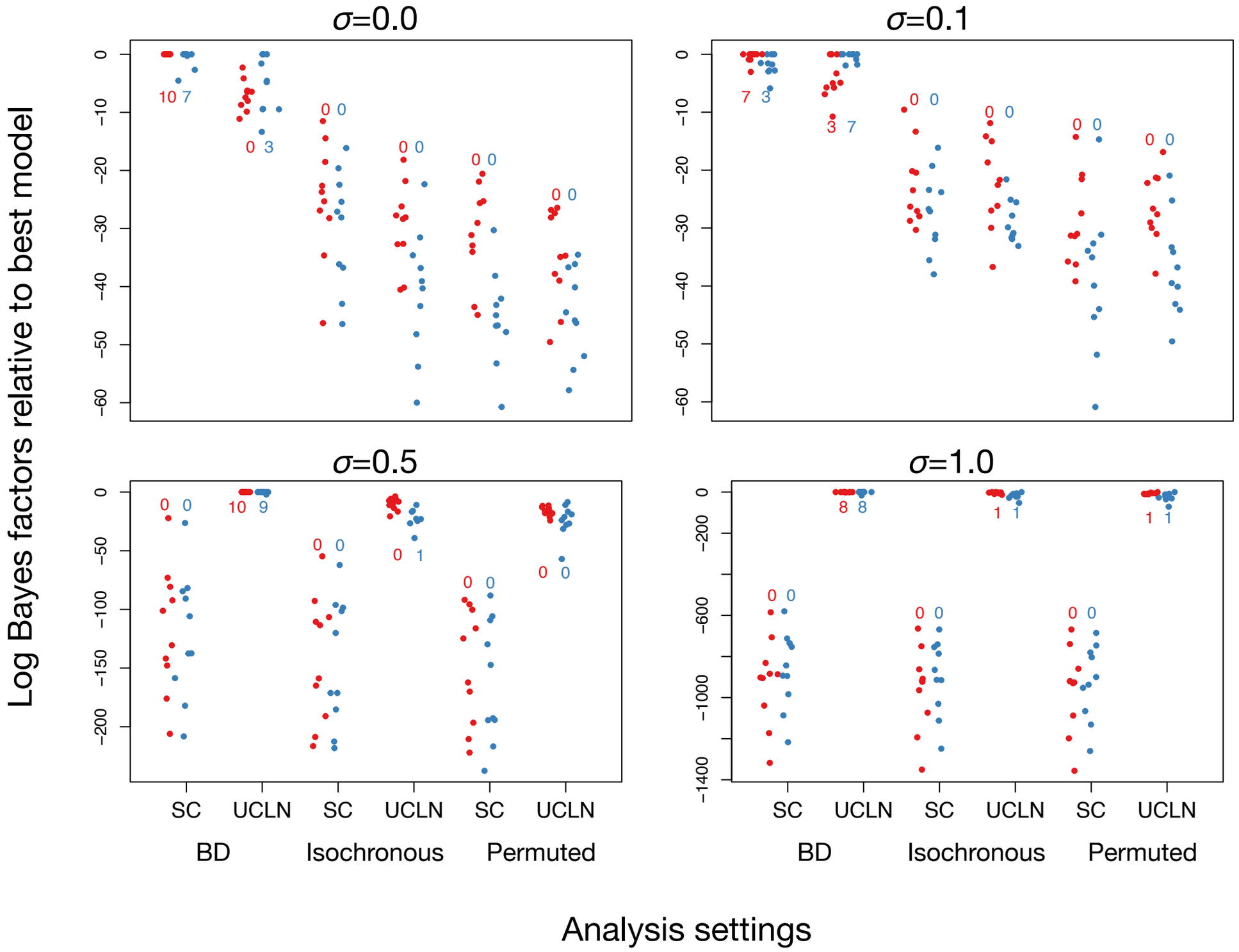
Fig. S1. Log Bayes factors of isochronous data simulated with a high substitution rate. Each panel shows the results for data sets simulated with a different degree of among-lineage rate variation, governed by the standard deviation σ of a lognormal distribution. The *x*-axis depicts six analysis settings, with two molecular clock models, strict clock (SC) and the uncorrelated relaxed clock with an underlying lognormal distribution (UCLN), and three settings for sampling times: generated under the birth-death process (BD), identical sampling times (Isochronous), and permuted (Permuted). The points have been jittered to facilitate visualization. The *y*-axis shows log Bayes factors relative to the best model. Red points correspond to estimates using generalized stepping-stone sampling and blue points correspond to estimates using nested sampling. We conducted 10 simulation replicates, with each replicate data set analysed under the six analysis settings and two marginal-likelihood estimators, such that stochastic error might cause differences in the preferred model. The number next to each cloud of points denotes the number of times (out of 10) that the corresponding model had the highest marginal likelihood with generalized stepping-stone sampling (red) and nested sampling (blue).

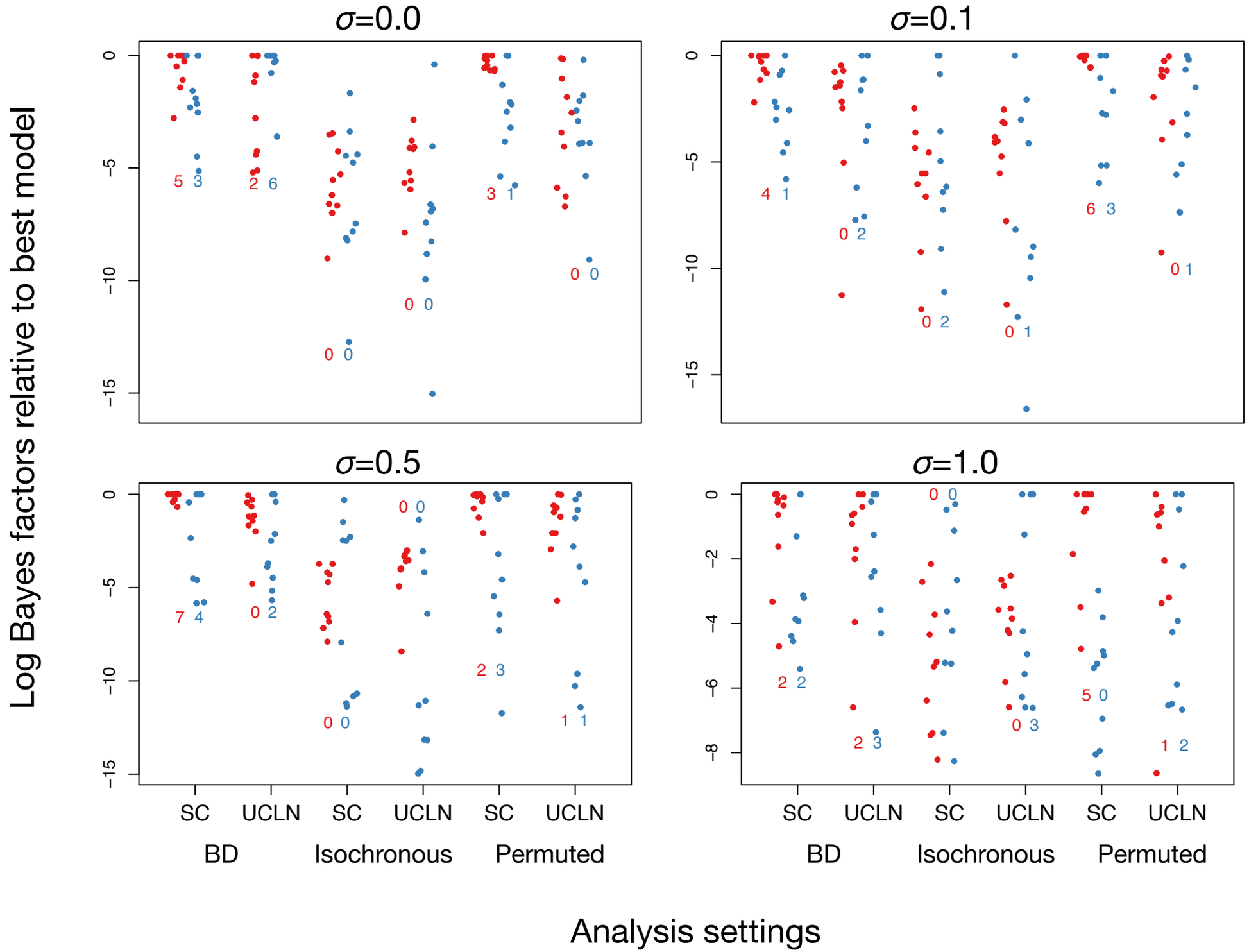
Fig. S2. Log Bayes factors of isochronous data simulated with a low substitution rate. Symbols and colours are the same as those in figure 1.

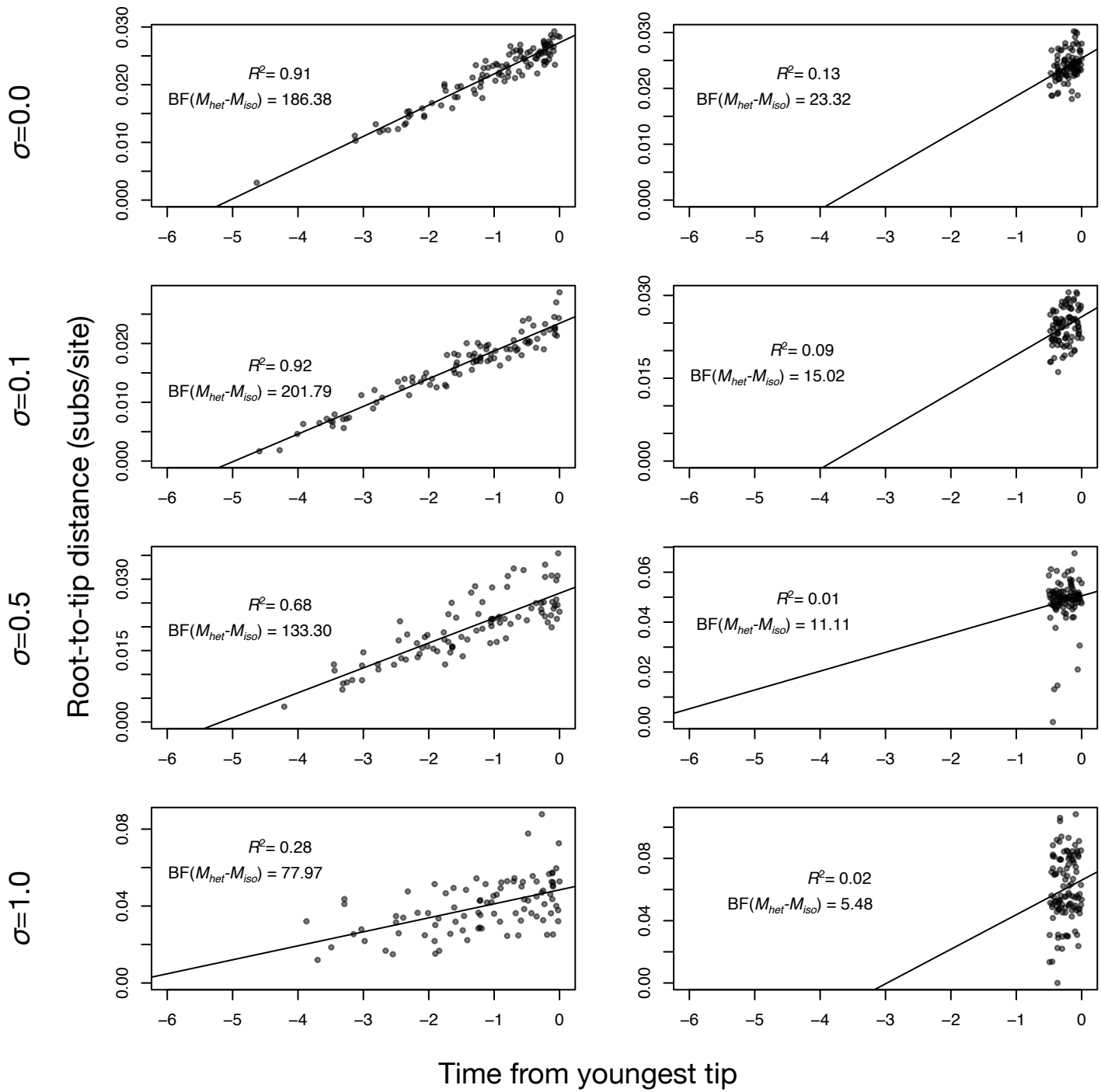
Fig. S3. Example of three phylogenetic trees used in simulations. Red dashed lines indicate the times of each of the tips and therefore represent the sampling process over time. All trees are simulated under a birth-death process with time of origin of 5, such that the sum of the tree height and the length of the stem branch leading to the root is always 5. In all trees, we set the birth rate λ=1.5, and become uninfectious rate δ=1, where $\delta = \mu + \psi$, where μ is the death rate and ψ is the sampling rate upon death. Thus, the population growth rate is constant and the same across all trees. The top tree assumes a constant sampling process and a wide sampling window (ψ=0.5 throughout the whole process), whereas in the second tree sampling starts after 4.5. Before this time the sampling rate, ψ_0, is zero. After 4.5 time units the sampling rate ψ_1 is 0.9 (and thus mu_1 = 0.1), resulting in a narrow sampling window. The bottom tree has samples drawn at a single point in time with a sampling probability at present, ρ, of 0.5 (and thus phi=0).

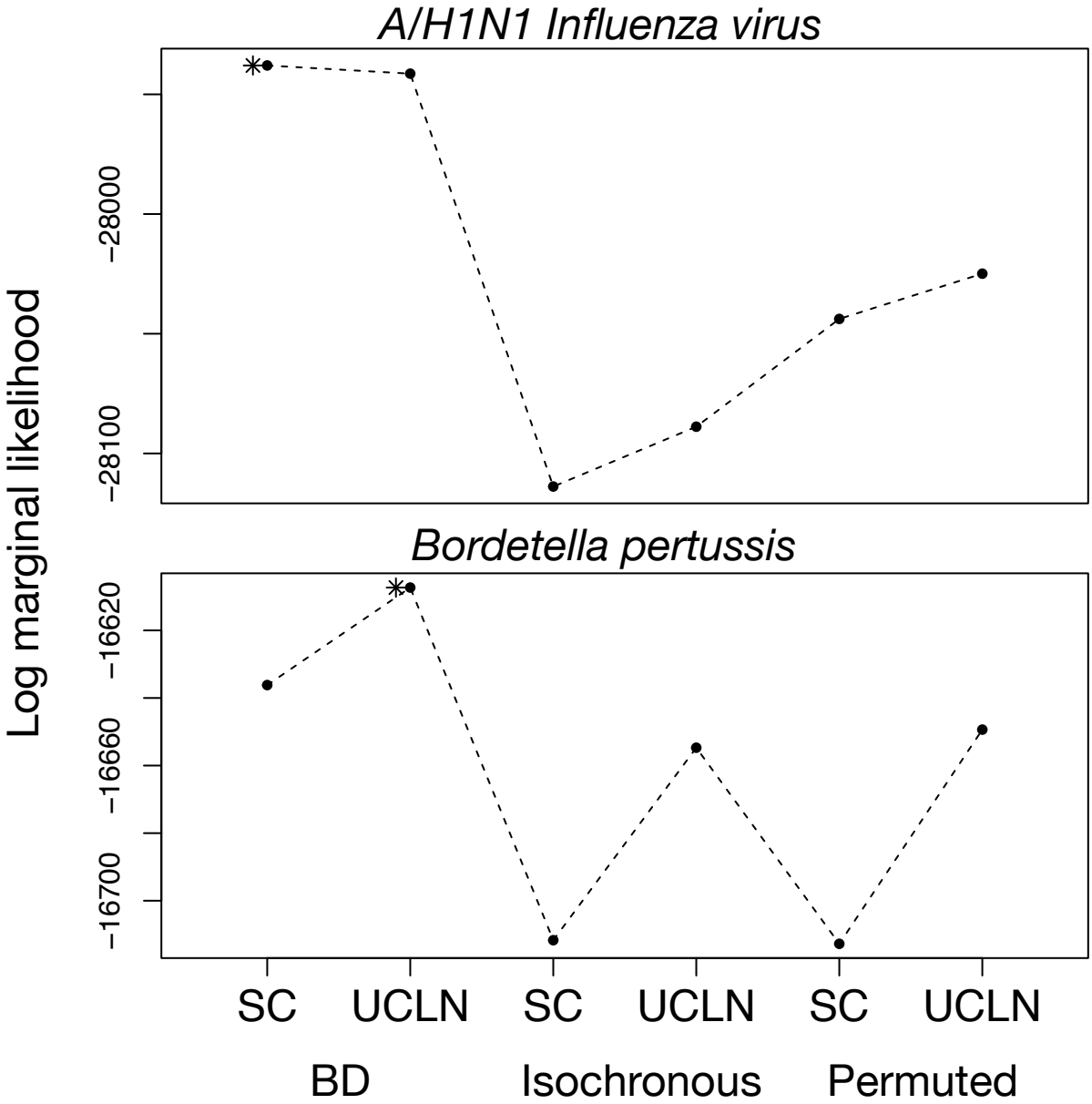

σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings

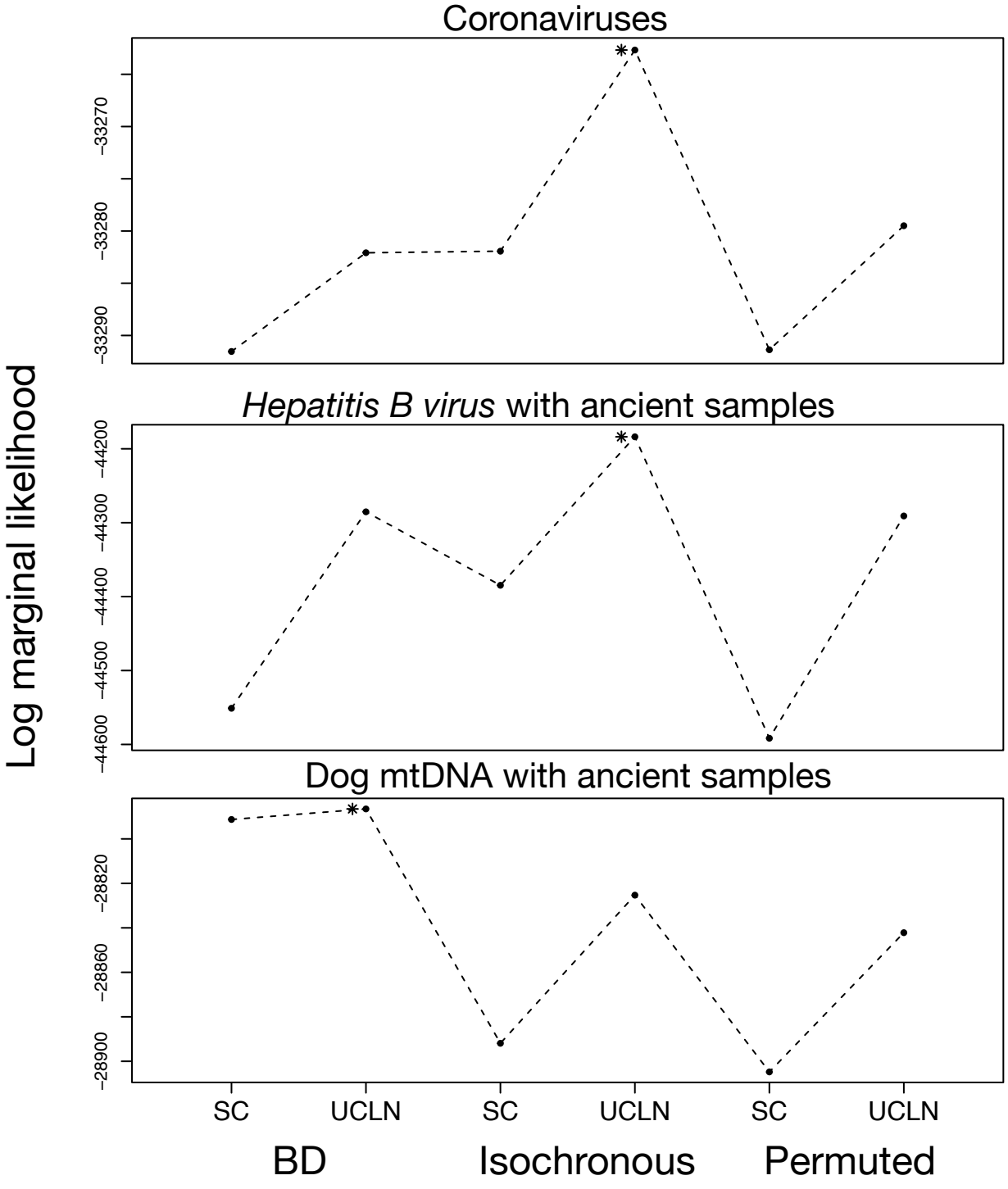

σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings

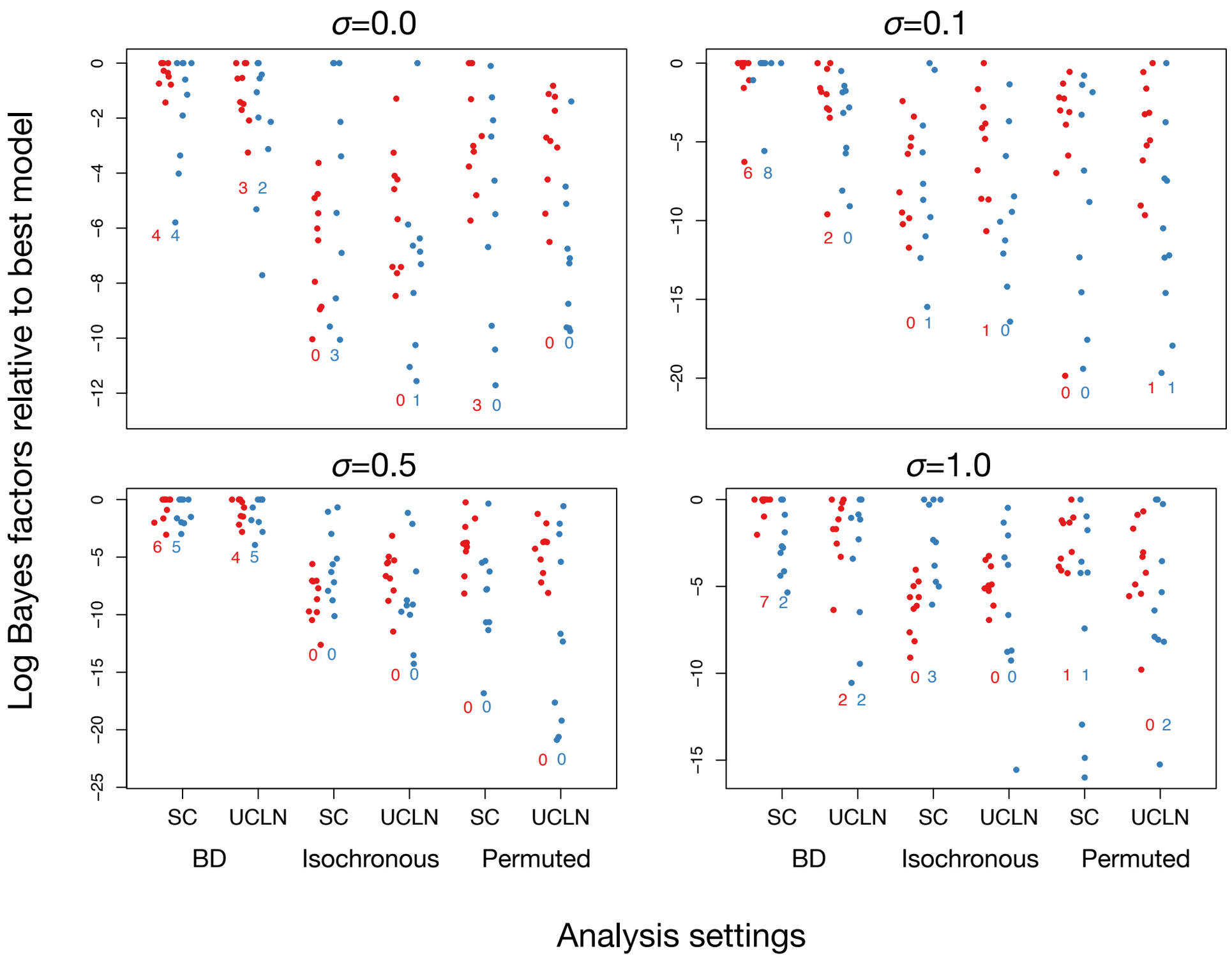

σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings

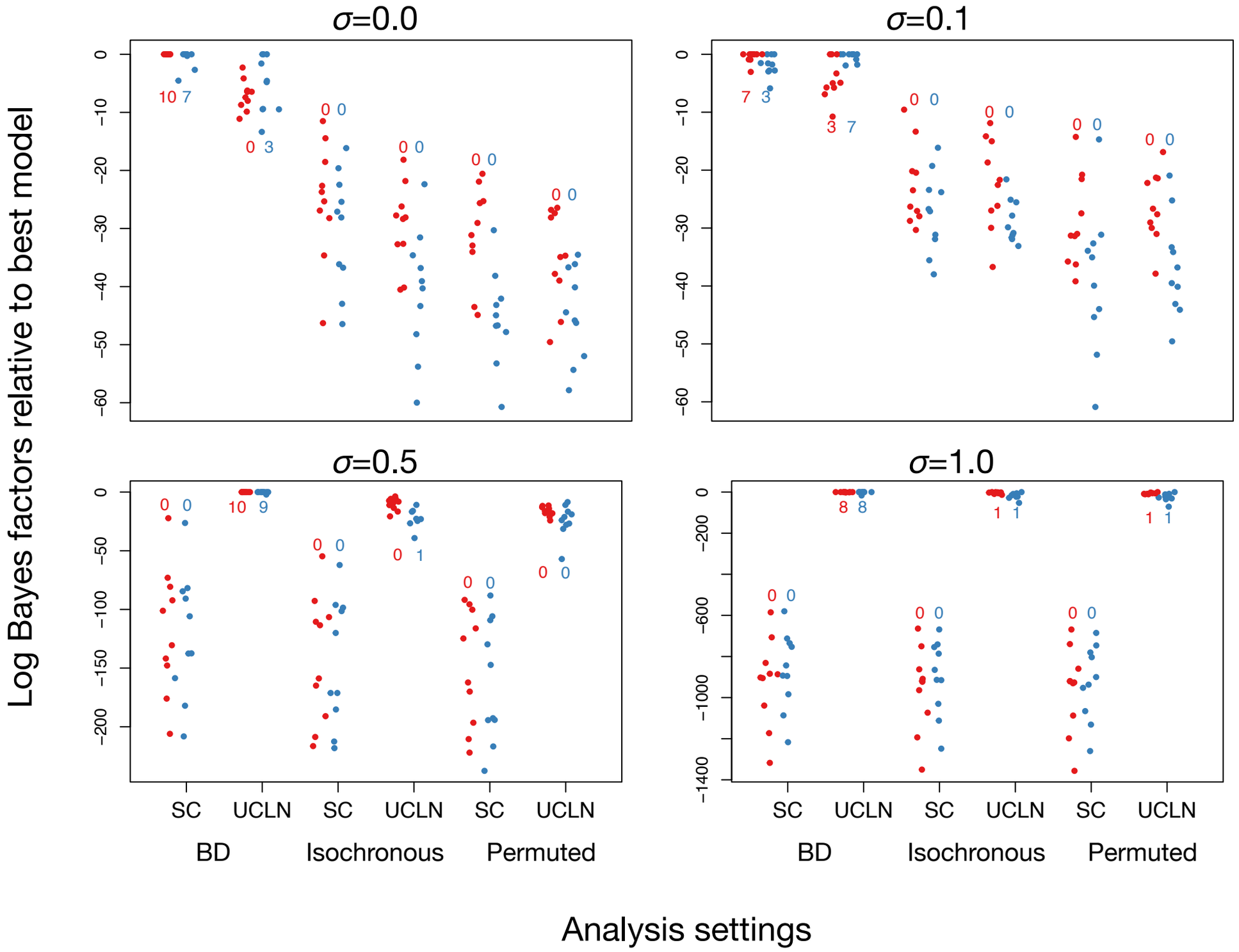

σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings

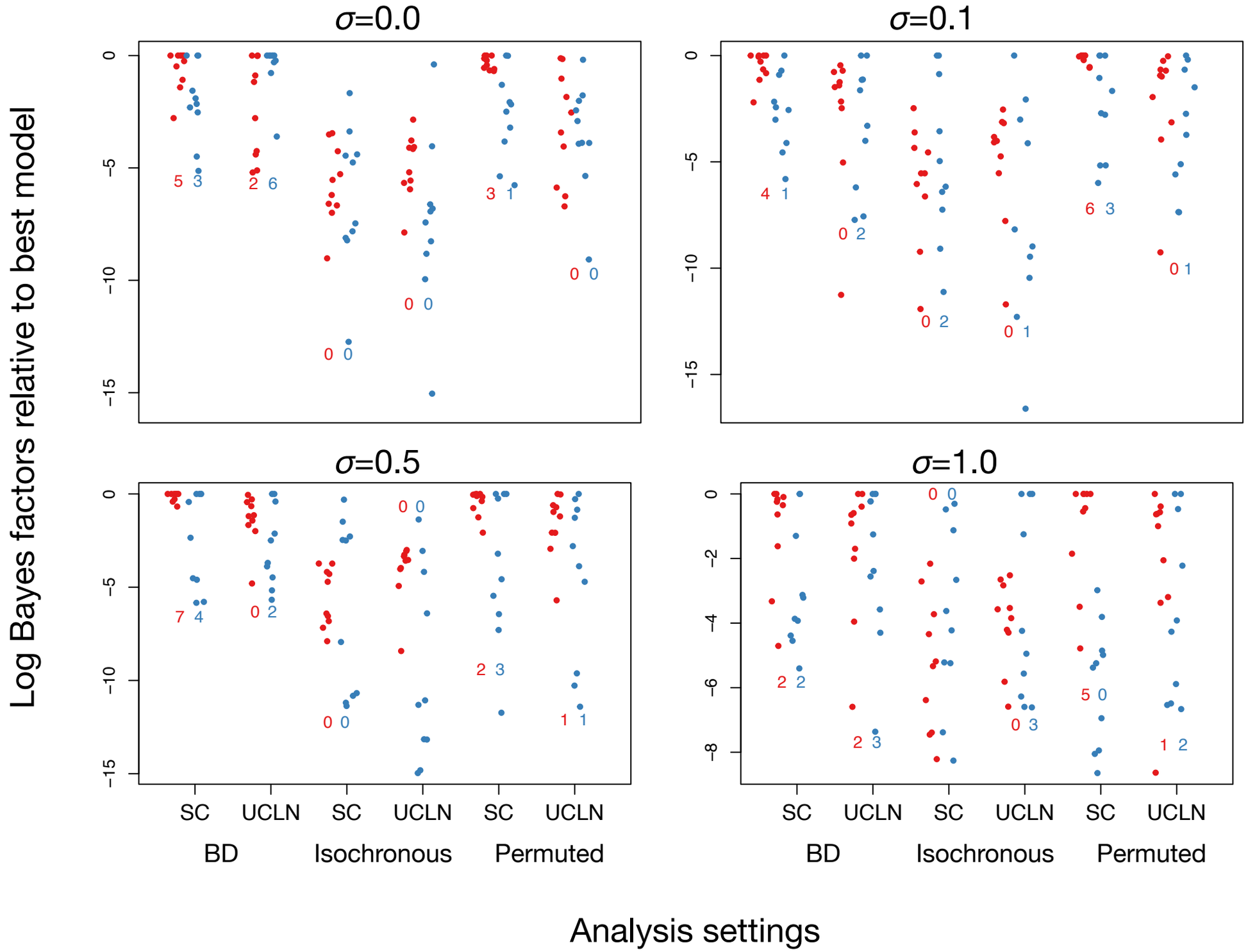

σ=0.0
σ=0.1
σ=0.5
σ=1.0
Root-to-tip distance (subs/site)
Time from youngest tip
R2= 0.91
BF(Mhet-Miso) = 186.38
R2= 0.13
BF(Mhet-Miso) = 23.32
R2= 0.92
BF(Mhet-Miso) = 201.79
R2= 0.09
BF(Mhet-Miso) = 15.02
R2= 0.68
BF(Mhet-Miso) = 133.30
R2= 0.01
BF(Mhet-Miso) = 11.11
R2= 0.28
BF(Mhet-Miso) = 77.97
R2= 0.02
BF(Mhet-Miso) = 5.48

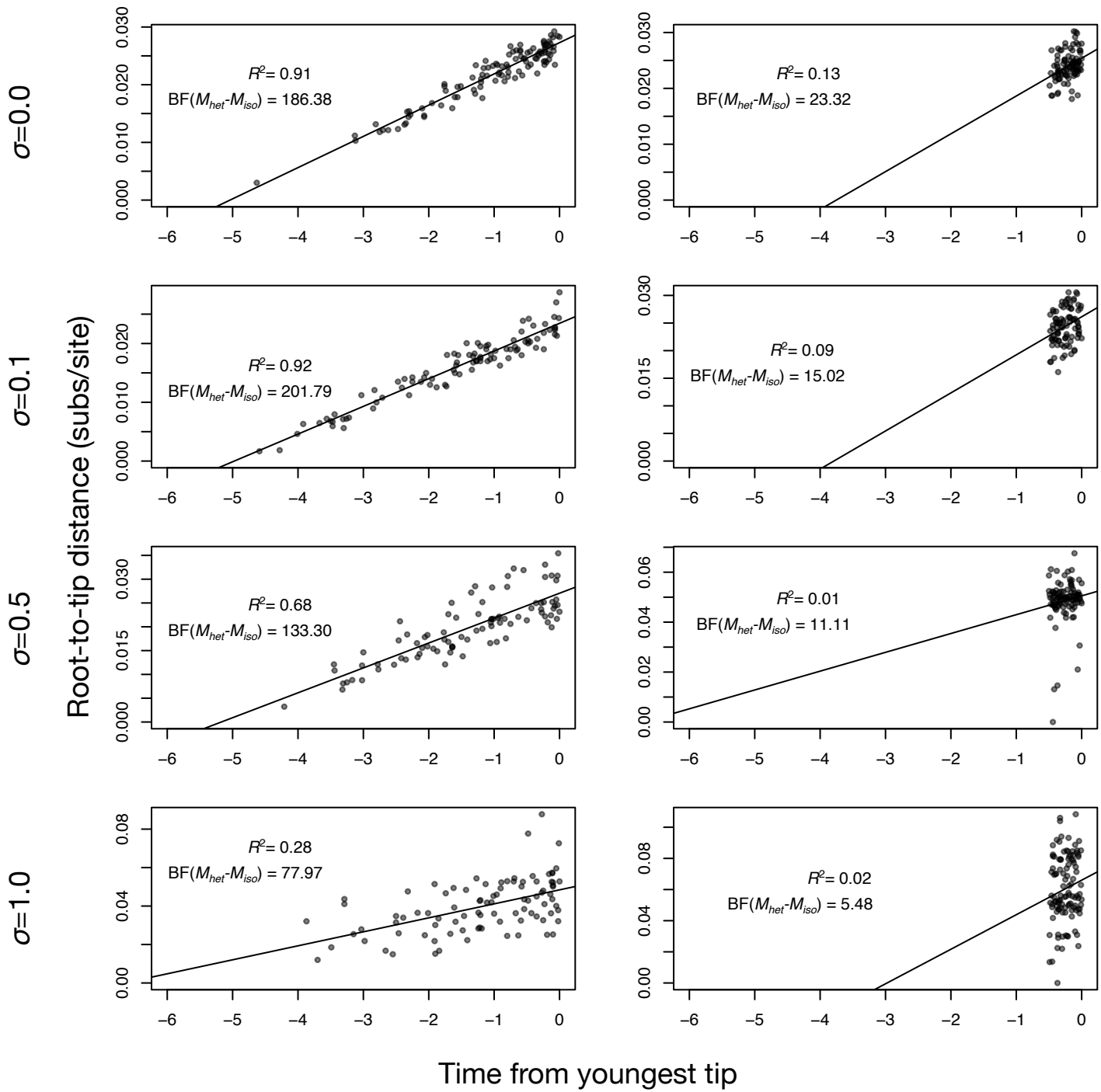

A/H1N1 Influenza virus
Bordetella pertussis
Log marginal likelihood
−28000
−28100
−16620
−16660
−16700
SC
UCLN
SC
UCLN
SC
UCLN
BD
Isochronous
Permuted

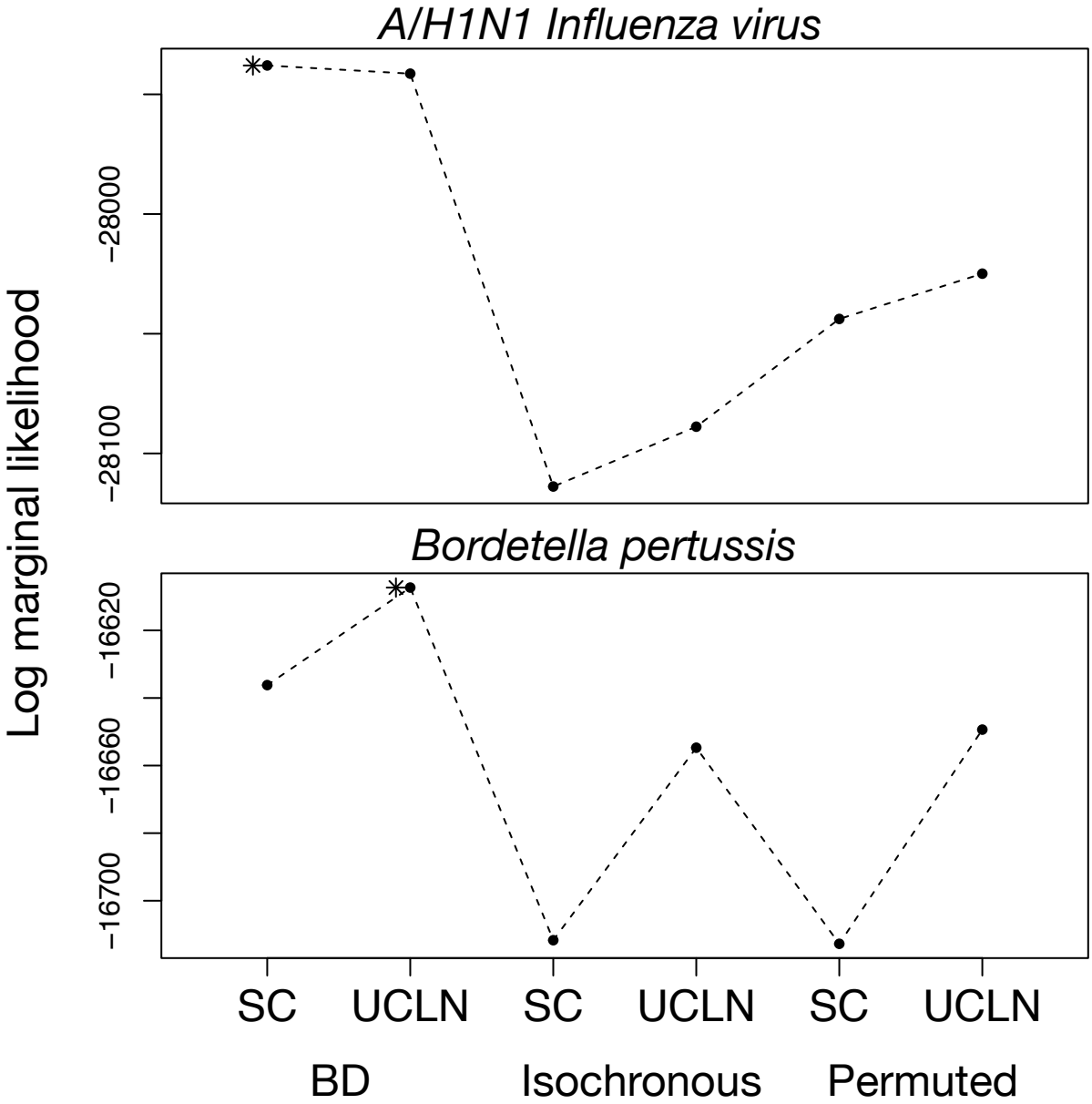

Coronaviruses
−33270
−33280
−33290
*
Hepatitis B virus with ancient samples
−44200
−44300
−44400
−44500
−44600
*
Dog mtDNA with ancient samples
−28820
−28860
−28900
*
Log marginal likelihood
SC
UCLN
SC
UCLN
SC
UCLN
BD
Isochronous
Permuted

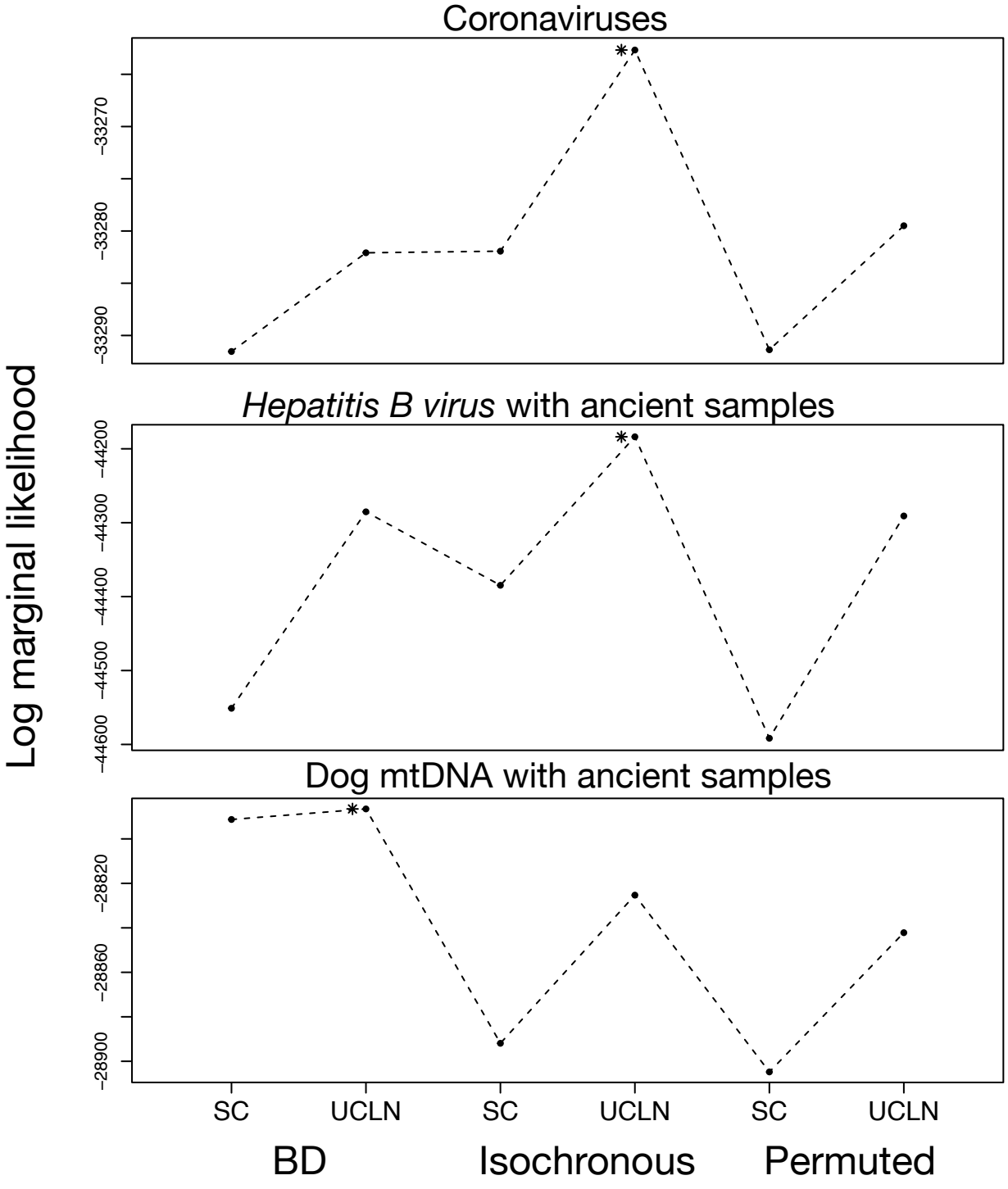

σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings


σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings


σ=0.0
σ=0.1
σ=0.5
σ=1.0
Log Bayes factors relative to best model
SC
UCLN
BD
Isochronous
Permuted
Analysis settings


σ=0.0
σ=0.1
σ=0.5
σ=1.0
Root-to-tip distance (subs/site)
Time from youngest tip
R2= 0.91
BF(Mhet-Miso) = 186.38
R2= 0.13
BF(Mhet-Miso) = 23.32
R2= 0.92
BF(Mhet-Miso) = 201.79
R2= 0.09
BF(Mhet-Miso) = 15.02
R2= 0.68
BF(Mhet-Miso) = 133.30
R2= 0.01
BF(Mhet-Miso) = 11.11
R2= 0.28
BF(Mhet-Miso) = 77.97
R2= 0.02
BF(Mhet-Miso) = 5.48


A/H1N1 Influenza virus
Bordetella pertussis
Log marginal likelihood
−28000
−28100
−16620
−16660
−16700
SC
UCLN
SC
UCLN
SC
UCLN
BD
Isochronous
Permuted


Coronaviruses
−33270
−33280
−33290
*
Hepatitis B virus with ancient samples
−44200
−44300
−44400
−44500
−44600
*
Dog mtDNA with ancient samples
−28820
−28860
−28900
*
Log marginal likelihood
SC
UCLN
SC
UCLN
SC
UCLN
BD
Isochronous
Permuted