# A new approach to Phylogenetic Model Comparison using MCMC

**Abstract**

High throughput sequencing has greatly improved our ability to investigate the evolutionary history of species using detailed demographic models. A popular approach for inferring parameters in these demographic models is by sampling genealogical histories at many short unlinked loci using a Markov Chain Monte Carlo algorithm. The use of explicit coalescent models by these methods makes them powerful for inferring demographic parameters, but they are limited in their ability to assess the fit of the inferred model to data. The purpose of this research is to examine a new approach, based on Relative Bayes Factors, for using genealogy samples to compare different evolutionary hypotheses.

**Demography Inference Methods and G-PhoCS**

Demography inference methods typically take in sequence data *X* from a collection of individuals originating from closely related populations and a parameterized demographic model *M*, and they infer values of parameters *Θ* in that model. Bayesian methods achieve this by assuming some prior distribution on the model parameters *P(Θ|M)* and sampling parameter values from an approximate posterior distribution *P(Θ|M, X*). Because the joint probability distribution *P(X , Θ|M)* cannot be efficiently computed, this task is often done by introducing local genealogies *G* to the model, such that the probability *P(X , G, Θ|M)* can be efficiently and accurately computed, and employing a Monte-Carlo Markov-Chain1 (MCMC) sampling algorithm for *G* and *Θ*, e.g. IMa 2, BP&P 3, and G-PhoCS4. Sampling by MCMC guarantees that *Θ & G* will be sampled from a probability distribution approximating the posterior–*P(Θ, G|X , M)*. From this distribution one can extract approximate posterior means and credible intervals for all demographic parameters.

G-PhoCS4 is one such Bayesian demography inference method. Given sequence data, i.e. a small collection of unphased genomes at tens of thousands of short unlinked neutrally evolving loci , and given a population phylogeny model, G-PhoCS infers parameters *Θ* such as population divergence times 𝜏, ancestral population sizes θ and rates of post-divergence gene flow *m*. This is accomplished using MCMC sampling of local genealogies jointly with model parameters according to an approximate posterior distribution for full Bayesian inference.

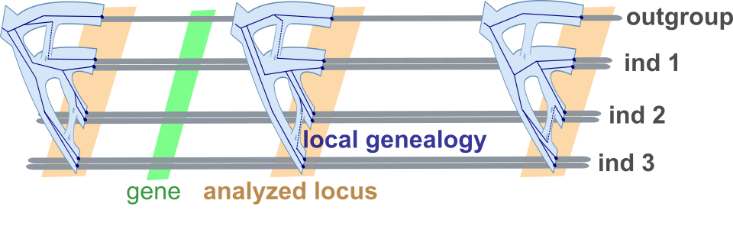


Fig 1 - G-PhoCS analyzing neutrally evolving loci

In each iteration G-PhoCS proposes a new instance of G, Θ and decides whether to accept or reject the new proposal based on the ratio of likelihoods of the current instance and the proposed instance. Formula [I] shows the likelihood calculation used by G-PhoCS –

[I]

In the above formula, is the Prior probability of model parameters to take current values. is the probability of the genealogy in locus given the model parameters. It is calculated under the Kingman Coalescent Model, with special regard to migration events. is the local data likelihood given the genealogy in locus . G-PhoCS assumes no recombination events within a locus, therefore the likelihood is a product across loci.

*Sequence data*

*Parameterss*

**The Model Comparison Problem**

More fundamental in the field of computational biology is the model comparison problem. ***The model comparison problem aims to compare the fit to sequence data between a collection of structural models***. Models in this case are phylogenetic topologies, also known as population or demographic models. An example Model-Comparison question would be – “*Given sequence data X of samples from relative populations, which of the candidate phylogenetic topologies {M} best fits the data?”*

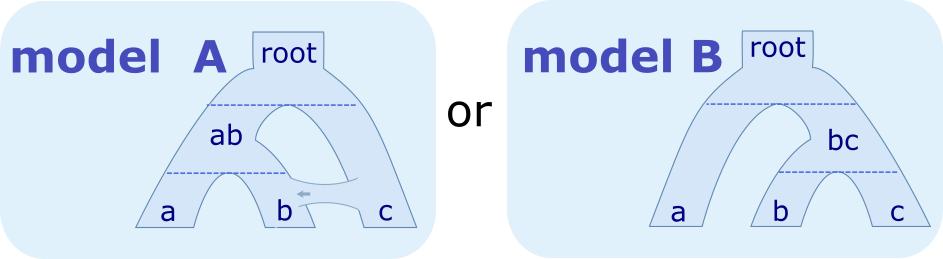


Fig 2 – An example of two feasible models MA & MB without Parameters Θ

The Model Comparison Problem makes a distinction between structural components of model M (tree topology, migration bands, parameter priors) and parameter values (specific migration rates, divergence times & population sizes). What model comparison aims to do is to compare different ‘model structures’.

Existing demography inference methods, as described in the introduction, don’t directly calculate likelihood - , with or without parameters . The consequence is that researchers are unable to efficiently test many model hypotheses to explain the sample DNA data and tackle the model-comparison problem.

In this study, building upon the G-PhoCS demography inference method and MCMC sampler, we intend to develop the theoretical framework and implement a method to compare multiple models M1, M2, …, Mi and their fit to the data, this without analytically calculating .

Bear in mind that in most cases researchers are interested only in qualitative claims about the structure of the model and not in qualitative claims about specific parameter values. To keep our method as general as possible, the comparison algorithm we present will receive no parameters *Θ* and will thus output a result pertaining only to the topology of the model. This will allow us to test *structural hypotheses* by integrating over parameter values.

Consider Figure 2 for example. In the figure is a comparison between models A & B. In model A, populations ‘a’ and ‘b’ are siblings and there is a constant migration rate across loci from population ‘c’ to ‘b’. In model A populations ‘b’ and ‘c’ are siblings (instead of ‘a’ and ‘b’) and there is no migration. The model-Comparison problems attempts to answer which model is more likely -

**An attempt – Standard Harmonic Mean**

One approach to the model-comparison problem is to directly estimate the likelihood of the two models, and . This is hard to analytically compute as X and M are only remotely related (via and ). One approach around this is the standard harmonic method. Defining as a joint random-variable of the genealogies and model parameters, we have –

[II]

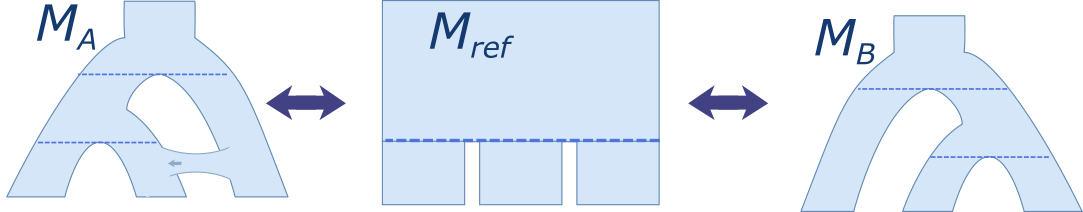
The last expression of formula [II] can be approximated using *n* MCMC samples of from the posterior distribution [G Θ|X, M] and calculating their mean. This approach fits naturally within the existing G-PhoCS framework of demography inference. However, since is a random variable with high variance (potentially unbounded when G is distributed according to- ), calculating its expectation is difficult and likely requires too large a number of samples 5.

**An improvement - Relative Bayes Factors**

In an attempt to improve statistical stability, we define and utilize in our calculations **a reference model - Mref.** A reference model is a demographic model which is a generalization (to be defined) of the models to be compared. Using Mref we compute the likelihood of every Mi relative to Mref. Then for every pair of models Mi and Mj we can compute their likelihood ratio.

[III]

The last expression of [III] can again be approximated using the existing G-PhoCS sampler. This approximation is superior to Harmonic-Mean as it has the potential for lower variance. Since we are free to choose a reference model, we can choose the one best suited for comparison of Mi and Mj. This allows us to finetune and minimize the variance of our calculation. A demonstration of this is to trivially choose , giving , with a variance of 0.



An example reference model Mref for comparing MA and MB

**Preliminary results**

To show viability of the Relative Bayes Factors theory we implemented a minimal model-comparison algorithm and testedit on a simple data set.

**Tests implementation and Setup**

For the tests we limited our choice of reference-models to models containing only a single ancestral population - , as shown in figure 3.

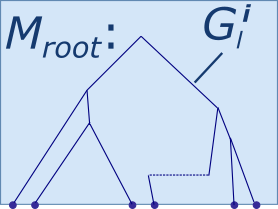


Fig 3 - The root reference Model Mroot

Formula [IV] is the genealogy likelihood calculation of Mroot under the Kingman Coalescent Model -

[IV]

We extended G-PhoCS to calculate sufficient statistics for Mroot, by having it emit in each iteration the aggregations of and across all loci. This allowed us to later calculate for any .

For now we completely omitted likelihood of parameter priors - , to be handled later in our research. We calculated instead an altered version of formula [III] –

[V]

For the experiments we generated two sequence data-sets, X0 and X1, under models <M0, Θ>and <M1, Θ> respectively. M0:= Mroot is the model containing a single population named *root* and no divergences or migrations. The parameter of the model Θ0 is population size θ­root = 0.001. M1 is a model containing a single divergence from population *root* into populations and with no migration. Parameters Θ1 are population times and divergence time as shown in figure 4.

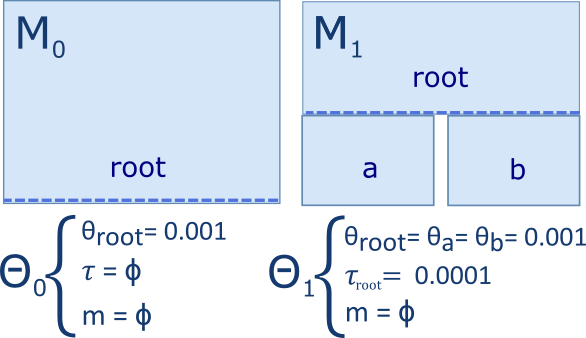
****

Fig 4 - Models used in preliminary experiments &

In both models we simulated genealogies over 16 samples (leaves). In M0 all samples belonged to the single population (root) and in M1 eight belonged to population *a* and eight to population *b*. Each data set contained 5,000 multiple sequence alignments of 16 sequences 1000 bases long.

Data was generated using standard tools; “ms” for generating data (genealogies or sequences) under the coalescence model (with recombination). “seq-gen” for simulating site-substitution models along genealogies6,7. Two more custom scripts were used for running a simulation given an “ms” command line file and for converting the "seq-gen” output file to a G-PhoCS sequence input file.

In each experiment G-PhoCS was run for 50,000 iterations. For each iteration was calculated using formula [IV]. As θ­root we chose to use the G-PhoCS sampled value of θ­root, regardless of the model used in the G-PhoCS analysis. For the expectation calculation, only the final 1000 samples were taken into account. This is done to avoid using data sampled before the MCMC sampler converges on the correct distribution - .

**Execution and Results**

We executed four simple experiments, each experiment characterized by its three inputs – sequence data, gphocs model and reference model - . The reference model always stayed Mroot and the sequence data and gphocs model alternated.

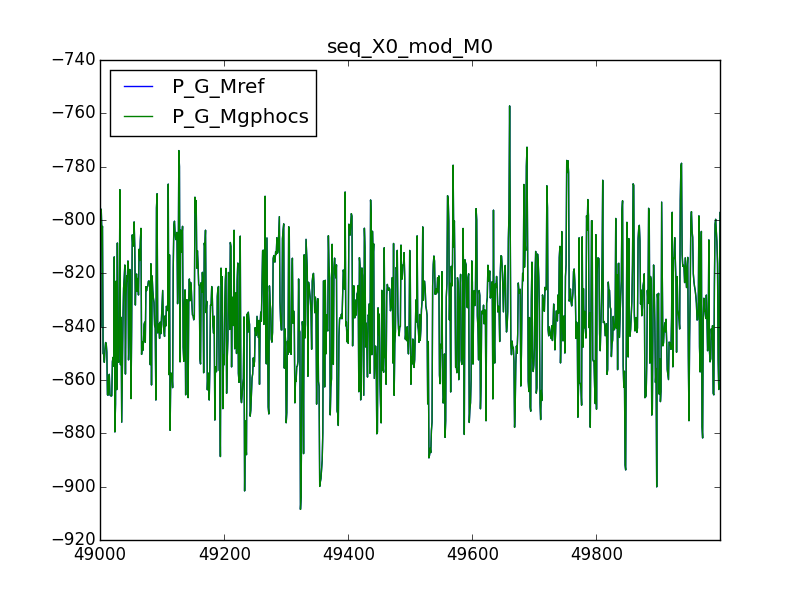
In the table below are logarithm results of the four combinations of experiments. Table columns and rows represent respectively the sequence data and phylogenetic model on which gphocs ran. For example the bottom right corner holds the result of experiment - :

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  | 11.55 | -57.12 |

The following is a summary of supporting data for each experiment:

* ***Experiment***  -

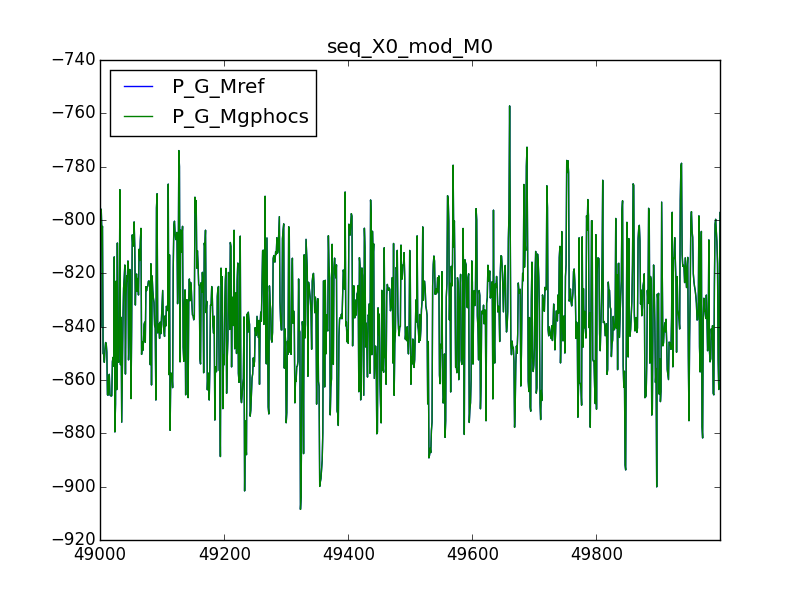
This first experiment is a rudimentary sanity check of the model-comparison algorithm. Since both the training model and reference model are equal to the root model Mroot, we expect -. Our result comes close to zero as we expected . The deviation from 0 is a result of the fact that we are computing in two different ways and may stem from rounding errors in various stages of the algorithm. We will revisit this discrepancy in future experiments.



A plot of and , sampled using X0, over G-PhoCS iterations. We see plots overlapping as expected since - . This constitutes a sanity check of the algorithm. Inaccuracies stem from the two different calculation methods.

* ***Experiment***  -

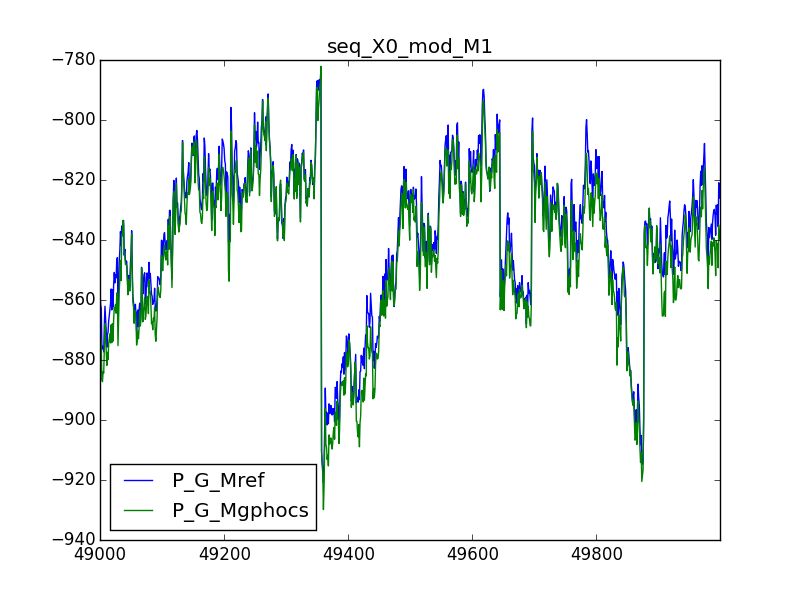
Again we expect - , as shown in the result . Contrary to prediction, we see the choice of data *Xi* effected the result. Further examination is due.



A plot of and , sampled using X1, over G-PhoCS iterations.

* ***Experiment***  -

Here we expect and our result is . The difference between likelihoods is small since during the G-PhoCS phase of the algorithm, the parameters of the training model M1 are fitted for sequence data *X0*, making it resemble Mroot by pushing the divergence time towards 0.

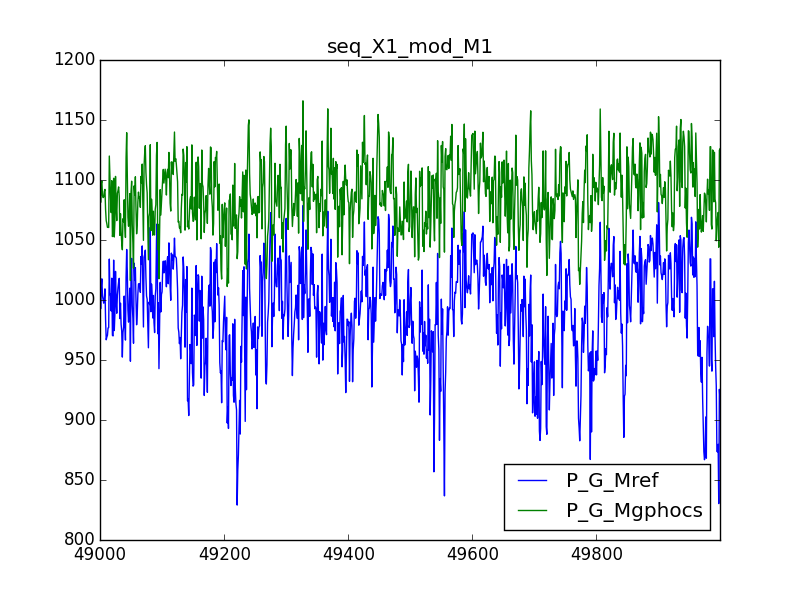


A plot of logP(G│M\_1 ) and logP(G│M\_root ), sampled using X0.

Estimated divergence time 𝜏root over gphocs iterations

* ***Experiment***  -

Here we expect and results confirm . The preference for M1 is strong since, contrary to the previous experiment, the G-PhoCS phase didn’t bring the gphocs and reference models closer to one another. Results are larger than 0 likely since they are samples of a density function, however this assertion needs to be reasserted by further research.



A plot of and , sampled using X1.

**Future Work**

**Main objectives**

Our research project has three parts – theoretical, algorithmic and experimental.

The objectives of the theoretical part are to figure out formal limitations for the reference model and to address the issue of models with different complexity, i.e. different sets of parameters and genealogies with or without migration events.

In the algorithmic part we will continually expand the capabilities of G-PhoCS and Model-Comparison algorithms. We’ll extended the sufficient statistics which we now output for Mroot, to allow likelihood calculations of more complex, possibly arbitrarily complex models. We’ll also solve the calculation inaccuracies, and perhaps add a reference-model selection phase.

In the experimental part we will continue work along the line we’ve set during the preliminary experiments phase, pushing forward each of the three dimensions of our experiments. We will generate and acquire data-sets of new populations and phylogenies, and train and compare them on new models.

As the purpose of the experiments is to test the sensitivity of our algorithm and to demonstrate its capabilities, we’re currently considering two strategies for assessing performance; an analytical measurements of variance of the estimate of RBF, and a comparison to the Harmonic-Mean estimator as a current baseline.

**1- Theoretical limitations of Mref**

The preliminary work has supplied for us a glimpse of possible results but it also accentuated the theoretical limitations of any reference model Mref. A primary goal of our research is then to formalize the “reference model” and its limitations: When calculating using formula [III], what are the implicit restrictions on your choice of Mref?

For example, one such restriction would be that should be well-defined. When calculating we are actually estimating a discrete average over samples from an estimated space - . For some reference models, this estimation breaks; If some genealogy G1 in M1 has zero probability in Mref, then P(G1|Mref)/P(G1|M) = 0. If this happens for many samples, our effective sample size is reduces and thus the estimation variance increases.A Criteria to meet this assumption is therefore something of the sort – *“Let Mref be any population model which gives non-zero probability to all G in M1”*. The laying down and formalization of these assumptions and criteria is an initial and primary step in our research.

**2- Handling Parameter Priors**

One area which was overlooked during the preliminary work, but is of theoretical and practical importance, is the difference between model parameters prior likelihoods. These are taken into account in our expectation estimation -

Different models M1, …, Mi as well as Mref are all structurally unique therefore they may contain different random variable parameters for population sizes, divergence times and migration rates (θ, 𝜏, m). Since priors are usually of the same gamma distribution, what matters however is mostly the number of different parameters |ΘM|, (|θ|, |𝜏|, |m|)M.

Currently, our preliminary algorithm completely ignores priors (formula [V] ). This is acceptable only under very limiting conditions such as |ΘM1| = |ΘM2|, because priors will reliantly cancel each other out. We must further explore and formalize our handling of priors: Under what assumptions may we ignore priors? How best to hand difference in priors between models without harming the correctness of model-comparison?

Another mile-stone in our work is to develop a theory and technique to map parameters from Mi to Mref in a way that satisfies [III], resulting in a legal Mref as described earlier. One idea for such a technique is to supplement models with “pseudo-parameters” – gamma distributed Random variables with no effect on the model - to equalize the number of parameters, then relying on the established algorithm for |ΘM1| = |ΘM2|. In our work we will drive to formalize and develop techniques for the handling of these parameters.

**3- Beyond the root reference model**

Our preliminary algorithm emitted only sufficient statistics *T* & *C* (explained in figure [IV]) and thus supported only the root reference model Mroot. For practical reasons our algorithm should aspire to support reference models as general as possible.

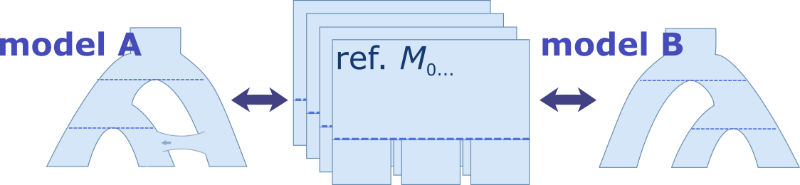
To do this we must redefine the sufficient statistics so that they will support more general reference models and then reimplement the algorithm to emit these statistics.

Depending on the scope of supported reference models, the statistics may range from a complete snapshot of genealogies across all loci in every iteration, through aggregating statistics for major events in the model such as population split, migration band start, similar to the one computed in our preliminary algorithm, all the way to a single precomputed likelihood value for some specific reference model.

The development and implementation of the data structures and algorithms required for this task will be the first and possibly largest ongoing challenges we will meet in our research.

**4- Choosing a reference model**

Once we have enabled calculating model-comparison using different reference models, we can now point our attention towards recommending the best one for a specific comparison. Formalizing and implementing an approach to finding the best Mref is necessary to make model-compare useful in real-world scenarios. The choice of a reference model can will influence the efficiency of estimation, making the difference between a quickly converging, high confidence, usable algorithm and an unreliable, very long-running algorithm (as described in the discussion about harmonic mean).



Choosing an optimal reference model

This is an advanced but rewarding part of our research. We will first describe a scoring function for the compatibility of Mref to the model-comparison question between Mi and Mj. Denote this function . We will rely on this function and suggest an algorithm to find an optimal Mref, one which maximizes.

**Milestones Timeline**

We’ve outlined our planned progress on each objectives -

* Handling Parameter Priors, 07/05/2016 – By this date we plan on defining the handling of parameter priors in *RBF*.
* Beyond the root reference model, 25/06/2016 – By this date we will implement a complete MC3 algorithm which gives the anthropology freedom in his choice of a reference model.
* Handling bias, 23/07/2016 – We expect the above implementation to suffer from a bias towards the hypotheses models. By this date we plan on minimizing this bias and improving the execution of the algorithm.
* Choose a reference model, 20/08/2016 – By this date we will develop another step in our Model-Compare algorithm, one which assists in choosing the best reference model for comparison of any two models.
* Theoretical limitations of Mref, 03/09/2016 – In this section of our work will address the theoretical limitations of a reference model.
* Putting down results, 15/10/2016 – This last month and a half will be dedicated to finalizing and writing down our work.

**References**

1. Hastings, W. K. Monte carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97–109 (1970).

2. Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 2785–90 (2007).

3. Yang, Z. & Rannala, B. Bayesian species delimitation using multilocus sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 9264–9 (2010).

4. Gronau, I., Hubisz, M. J., Gulko, B., Danko, C. G. & Siepel, A. Bayesian inference of ancient human demography from individual genome sequences. *Nat. Genet.* **43,** 1031–1034 (2011).

5. Newton, M. a & Raftery, A. E. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. Ser. B* **56,** 3–48 (1994).

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

7. Hudson, R. R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinforma. Appl. NOTE* **18,** 337–338 (2002).