Current Zoology 61 (5): 854–865, 2015

The BPP program for species tree estimation and species delimitation

Ziheng YANG*

[1] Center for Computational Genomics, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[2] Department of Genetics, Evolution and Environment, University College London, Darwin Building, Gower Street, London WC1E 6BT, England

Abstract This paper provides an overview and a tutorial of the BPP program, which is a Bayesian MCMC program for analyzing multi-locus genomic sequence data under the multispecies coalescent model. An example dataset of five nuclear loci from the East Asian brown frogs is used to illustrate four different analyses, including estimation of species divergence times and population size parameters under the multispecies coalescent model on a fixed species phylogeny (A00), species tree estimation when the assignment and species delimitation are fixed (A01), species delimitation using a fixed guide tree (A10), and joint species delimitation and species-tree estimation or unguided species delimitation (A11). For the joint analysis (A11), two new priors are introduced, which assign uniform probabilities for the different numbers of delimited species, which may be useful when assignment, species delimitation, and species phylogeny are all inferred in one joint analysis. The paper ends with a discussion of the assumptions, the strengths and weaknesses of the BPP analysis [*Current Zoology* 61 (5): 854–865, 2015].

Keywords BPP, MCMC, Multispecies coalescent, Species delimitation, Species tree

1 Introduction

1.1 Overview of bpp

BPP (for Bayesian Phylogenetics and Phylogeography) is a Bayesian Markov chain Monte Carlo (MCMC) program for analyzing DNA sequence alignments under the multispecies coalescent model (MSC) (Rannala and Yang, 2003; see also Takahata et al., 1995; Yang, 2002). The MSC model lies at the interface of molecular phylogenetics and population genetics. Compared with traditional phylogenetic analysis, which assumes that the same tree underlies all gene loci, the MSC accounts for the coalescent process in both the modern and ancestral species and the resultant gene tree-species tree conflicts. Thus a reliable estimation of the species phylogeny is possible even if the information at every locus is weak so that the gene tree is highly uncertain (Heled and Drummond, 2010). Edwards (2009) has argued that species tree estimation under the MSC represents a paradigm shift in molecular phylogenetics. Compared with traditional population genetics models (in particular models of population subdivision), MSC accounts for the phylogenetic history of the species or populations. For many datasets, this is more realistic than an equilibrium model of population subdivision and migration. The MSC provides a natural framework for addressing a number of important problems in evolutionary biology, such as species delimitation (Yang and Rannala, 2010; Rannala and Yang, 2013), species tree estimation (Edwards et al., 2007; Liu and Pearl, 2007; Heled and Drummond, 2010), and detection of hybridization and contamination. See Fujita et al. (2012) and Carstens et al. (2013) for reviews on species delimitation methods using genetic sequence data. A critical assessment of the strengths and weaknesses of those methods is provided by Rannala (2015). Liu et al. (2015) reviewed methods for species tree estimation in the presence of conflicting gene trees.

The MSC has been extended to account for migration between populations (Hey, 2010) and recombination along the sequence (Hobolth et al., 2007). Here in this paper we focus on the basic MSC model. The model includes two types of parameters: the species divergence times (τs) and the population size parameters for both modern and ancestral species (θs). Here we use the example of East Asian brown frogs in the *Rana chensinensis* species complex to illustrate the model (Fig. 1). There are four populations or species: represented as K, C, L, and H. If the phylogeny is ((K, C), (L, H)), as shown in Fig. 2A, there will be three species diver-

Received Mar. 13, 2015; accepted Apr. 30, 2015.
* Corresponding author. E-mail: z.yang@ucl.ac.uk

gence times (τ_{KC}, τ_{LH}, and τ_{KCLH}) and seven population sizes (θ_K, θ_C, θ_L, θ_H, θ_{KC}, θ_{LH}, and θ_{KCLH}). In general, for a species tree of s species, there will be $s - 1$ divergence time parameters and at most $2s - 1$ population size parameters in the MSC model. Note that two sequences are needed to calculate a distance, so that if there is no or only one sequence from a population at every locus, it will not be possible to estimate the θ parameter for that population. Both τs and θs are measured by the sequence distance or the expected number of mutations per site. In particular, $\theta = 4N\mu$, where N is the (effective) population size and μ is the mutation rate per site per generation, so that $\theta_K = 0.002$ in figure 2b means that two sequences taken at random from population K have 2 differences per kilobase. Note that it takes on average $2N$ generations for two sequences taken at random from a population to coalesce (to find their common ancestor), so that the distance between the two sequences is $2N \times \mu \times 2 = \theta$.

Four types of analysis are possible by BPP, referred to in this tutorial as A00, A01, A10, and A11. These are specified using two variables (switches) in the control file: `speciesdelimitation` and `speciestree`. The four analyses are summarized in table 1.

Note that analysis A00 is a within-model inference, and its objective is to generate the posterior distribution of the parameters (θs and τs) under the MSC model. The MCMC algorithm implemented in the BPP program for this inference appears to be fairly efficient and has been applied to genomic datasets consisting of ~50,000 loci (Burgess and Yang, 2008). The other three analyses (A01, A10, and A11) are transmodel inferences (in the terminology of Green, 2003), in which the Markov chain moves between different models. Each of those models is an instance of the MSC model, but the number and nature of the species and the species phylogeny

Fig. 1 Brown frogs from the C and L clades of *R. chensinensis* and a map showing the geographical distributions of all four clades of East Asia brown frogs

Clades C and L of *R. chensinensis*, clade K (*R. kukunoris*) and clade H (*R. huanrensis*). Photos courtesy of Dr Hui Zhao, Institute of Biology, Chinese Academy of Sciences, Chengdu, China.

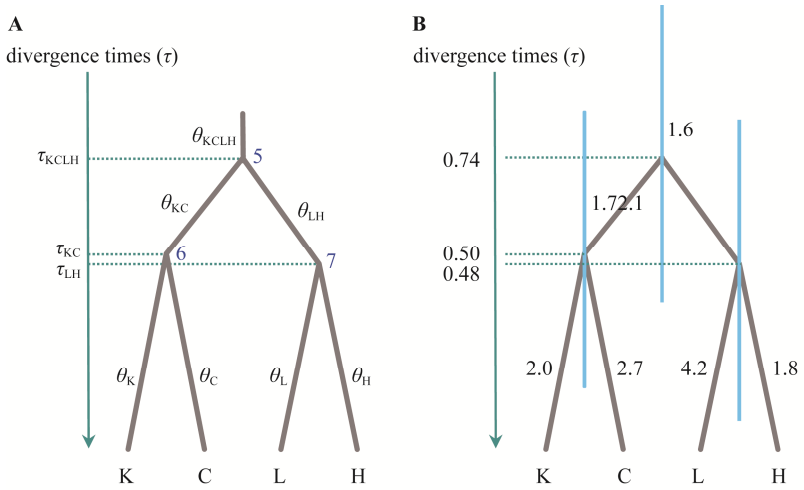

Fig. 2 Analysis A00

A. A species tree for four brown frog species/populations K, C, L, and H, illustrating the parameters in the multispecies coalescent model. Those include three species divergence time parameters (τ) for the three ancestral nodes, 5 (KCLH), 6 (KC), and 7 (LH), and seven population size parameters (θs) for the seven populations on the tree. **B**. Estimates (posterior means) of the parameters obtained from analyzing the sequence data at five nuclear loci. Both τs and θs are measured by the expected number of mutations per kilobase. The priors used in the analysis are $\theta \sim$ G(2, 1000) for all populations and $\tau_{KCLH} \sim$ G(2, 2000) for the root age. The node bars represent the 95% HPD intervals for divergence times. The tree is drawn with FIGTREE using the BPP output.

Table 1 Four analyses implemented in bpp and illustrated in the tutorial (A00, A01, A10, and A11)

Speciesdelimitation	Speciestree	
	0	1
0	**A00**. Estimation of parameters under the multispecies coalescent model (τs and θs) when the species phylogeny is given (Yang, 2002; Rannala and Yang, 2003).	**A01**. Inference of the species tree when the assignment and delimitation are given (Rannala and Yang, ms. in preparation)
1	**A10**. Species delimitation using a fixed guide tree (Yang and Rannala, 2010; Rannala and Yang, 2013).	**A11**. Joint species delimitation and species-tree inference or unguided species delimitation (Yang and Rannala, 2014).

may differ among those models. The main objective of the transmodel inference is the calculation of the posterior probabilities for the different models. We have found cases in which the Markov chain mixes poorly in the transmodel algorithms. The mixing problem is discussed later.

1.2 Running the bpp program

Detailed information about downloading and compiling the BPP program is provided at the program web site and in the program manual (bppDOC.pdf). Here we go through the basic steps, but the manual should be consulted for more details. The manual also explains the format of the data files, the screen output, as well as the output files.

Download BPP from the web site http://abacus.gene.ucl.ac.uk/software/. The current version is 3.1, which we will use here. The archive includes Windows executables and ANSI C source files. For UNIX/LINUX or MAC OSX, you need to compile the program first. For example, the following command uses the compiler `gcc` to generate the executable `bpp`.

```
gcc -o bpp -O3 bpp.c tools.c -lm
```

In the tutorial below, we will use a dataset of five nuclear loci from the East Asia brown frogs (Zhou et al., 2012). The sequence alignment (`frogs.txt`) and the Imap (`frogs.Imap.txt`) files, as well as the control files for the four analyses, are in the folder `frogs` in the BPP release.

We will run BPP from the command line, although you may use the BPPX interface (written by Bo Xu). If you have not used the command line before, please work through one of the following tutorials first:

http://abacus.gene.ucl.ac.uk/ziheng/CommandLine.Windows.pdf;

http://abacus.gene.ucl.ac.uk/ziheng/CommandLine.MACosx.pdf.

We will run each analysis twice in two folders, `r1` and `r2` inside the frogs folder. Start two command-line terminals. Then change directory to `r1` (or `r2`), and run the program as follows.

On Windows	On LINUX/UNIX/MAC OSX
`cd frogs\r1`	`cd frogs/r1`
`..\..\bpp ..\A00.bpp.ctl`	`../../bpp ../A00.bpp.ctl`

Here `A00.bpp.ctl` (in the frogs folder) is the control file for analysis A00 (Fig. 3). To run the other three analyses, replace A00 with A01, A10, or A11. Note that in the control file (Fig. 3), the data file is specified as `../frogs.txt` instead of `frogs.txt`, because the file is in the frogs folder while we run BPP in the `frogs/r1` folder.

The run will produce an MCMC sample file (`mcmc.txt`), which is summarized by BPP. The output (`out.txt`) should be self-explanatory, but see the manual for detailed explanations. Running the same analysis multiple times allows us to confirm that the results are stable across runs. You may also merge the two samples into one and summarize the combined sample: Append one file to the end of the other (and remove the header line of the second file if it exists). Then run BPP with `print = -1`.

2 The Example Dataset of East Asian Brown Frogs

We use the five nuclear loci from East Asian brown frogs in the *Rana chensinensis* species complex (Zhou et al., 2012). Three morphologically recognized species exist in this group: *R. chensinensis* (clades C and L), *R. kukunoris* (K) and *R. huanrensis* (H). *R. chensinensis* has a widespread distribution in northern China. *R. kukunoris* occurs on the eastern edge of the Qinghai-Tibetan Plateau, while *R. huanrensis* has a limited distribution in Northeast China and Korea (Fig. 1). Those geographical areas have very different climates and correspond to different ecological habitats. Divergences of those species have been hypothesized to coincide with tectonic and climatic changes associated with the uplifting of the Qinghai-Tibetan Plateau (Zhou et al., 2012).

Zhou et al. (2012) conducted a phylogenetic analysis of an extensive sample of a mitochondrial locus (cyt *b*).

```
          seed = -1
       seqfile = ../frogs.txt
      Imapfile = ../frogs.Imap.txt
       outfile = out.txt
      mcmcfile = mcmc.txt

  speciesdelimitation = 0 * fixed species tree
* speciesdelimitation = 1 0 2    * speciesdelimitation algorithm0 and finetune(e)
* speciesdelimitation = 1 1 2 1 * speciesdelimitation algorithm1 finetune (a m)
*         speciestree = 1        * speciesdelimitation algorithm1 finetune (a m)
*   speciesmodelprior = 1  * 0: uniform LH; 1:uniform rooted trees; 2: uniformSLH; 3: uniformSRooted
  species&tree = 4  K  C  L  H
                    9  7 14  2
                   ((K, C), (L, H));

       usedata = 1  * 0: no data (prior); 1:seq like
         nloci = 5  * number of data sets in seqfile
     cleandata = 1    * remove sites with ambiguity data (1:yes, 0:no)?
    thetaprior = 2 1000   # gamma(a, b) for theta
      tauprior = 2 2000 1  # gamma(a, b) for root tau & Dirichlet(a) for other tau's
      finetune = 1: 5 0.001 0.001 0.001 0.3 0.33 1.0  # for GBtj, GBspr, theta, tau, mix

        print = 1 0 0 0   * MCMC samples, locusrate, hereditystalars, Genetrees
       burnin = 8000
     sampfreq = 2
      nsample = 100000
```

Fig. 3 Control file A00.bpp.ctl for analysis A00 (with `speciesdelimitation = 0` and `speciestree = 0`)
This is set up so that the BPP program is launched in the folder frogs\r1, while the sequence and Imap files are in the folder frogs. The total number of MCMC iterations is `burnin` + `nsample` × `sampfreq` = 208 000. Note that lines starting with an asterisk are comments and the default values of `speciesdelimitation` and `speciestree` are 0.

The maximum likelihood tree corresponds very well with the geographical distribution of the species with two exceptions. First, the analysis identified four major clades instead of three species, with *R. chensinensis* split into clade C from the type locality and clade L from the Loess Plateau. The two clades may represent two distinct species. Second, some individuals of *R. chensinensis* from the western Qinling Mountains form a subclade within *R. kukunoris* on the mitochondrial tree, while morphology and nuclear loci suggest they belong to the C clade. This appears to be a case of mitochondrial introgression from *R. kukunoris* into *R. chensinensis*.

In the tutorial below, we use BPP to analyze the five nuclear loci for a subset of the samples sequenced by Zhou et al. (2012) to represent the major lineages on the mitochondrial tree. The loci are RAG-2 (440 bp, 28 sequences), Tyr (455 bp, 28 sequences), BDNF (457 bp, 30 sequences), POMC (285 bp, 24 sequences) and PG (489 bps, 21 sequences). The sequences are assigned to the four populations K, C, L and H. Three *R. chensinensis* samples apparently represent mitochondrial introgression and are thus assigned back to clade C in our analysis of the nuclear loci (Zhou et al., 2012).

Each analysis here takes three input files: the control file (e.g., `A00.bpp.ctl`), the sequence alignment file (`frogs.txt`) and the Imap file (`frogs.Imap.txt`), with the latter two files specified in the control file. The sequence alignments are in the PHYLIP/PAML format, with one alignment following the other, all in one file. Alignment gaps and ambiguity nucleotides are either deleted before analysis (`cleandata = 0`) or used in the likelihood calculation (see Yang, 2014, pp. 111-2).

The Imap file assigns individuals or sequences to the populations. In the sequence data file, each sequence name has a tag (indicated by ^) which is interpreted as an individual ID and used in the Imap file to assign the sequence to a population. For example, the sequence name ^kiz1375 in the sequence data file has the tag kiz1375, which is used in the Imap file to assign specimen kiz1375 to population C. Thus through the Imap file, each sequence is assigned to a population. All models currently implemented in BPP use the population ID for each sequence but ignore the individual ID; for example, information in linkage disequilibrium among loci is ignored. One could tag each sequence by the population ID to avoid the need for the Imap file. However, the current setup allows one to change the assignments easily without editing the large sequence data file.

3 The Tutorial

In this tutorial, we conduct all four analyses listed in Table 1: A00, A01, A10 and A11. We run each analysis

at least twice (in `r1` and `r2`, inside the `frogs`/ folder). With 208,000 (= `burnin` + `nsample` × `sampfreq`) iterations, each run took about 10 minutes on a laptop. If the results look too different between runs, we re-run the program using a larger number of samples (`nsample`) and/or larger sampling frequency (`sampfreq`) (Fig. 3).

There are standard tools for diagnosing convergence and mixing problems of MCMC algorithms (Robert and Casella, 2004, pp.459–510; Yang, 2014, pp. 226–244). For running BPP, our experience suggests that running the same analysis multiple times appears to be the most effective method to guarantee the reliability of the results. There are no hard rules for deciding how large a difference between runs is too large, so common sense is advised. The main objective of analyses A01, A10, and A11 is to calculate the posterior probabilities of models. It is advisable to calculate those to the percentage point (e.g., 71%), but it may not be necessary to calculate them to the first decimal point (e.g., 70.9%). Similarly analysis A00 generates the posterior distribution of the parameters under the MSC. A 1% or 5% relative error in the posterior means or in the posterior interval limits may be precise enough; for example, 0.0020, as an estimate of the posterior mean of θ_K accurate to the fourth decimal point, may be precise enough and there is no need to calculate it to 0.00196 (Fig. 2B). Such choices of course depend on the computing resources available, the absolute running time, etc. In this regard, note that the variance of the estimate of the posterior model probability based on an MCMC sample of size N is $P(1 - P)/(NE)$, where P is the true posterior model probability, and E is the efficiency of the MCMC sample (see, e.g., Yang, 2014, p.214). Note that the estimate based on an independent sample has the variance $P(1 - P)/N$ and such an estimate has efficiency E = 100%. Similarly, the variance of the posterior mean of a parameter based on an MCMC sample of size N is $v/(NE)$, where v/N is the variance based on an independent sample. In both cases of calculating posterior model probabilities (A01, A10, and A11) and estimating model parameters (A00), the variance is proportional to $1/N$. Thus to reduce the standard error of the estimate by a half one has to increase the MCMC sample size by four folds; or to increase the number of significant digits by one (or to increase the estimation precision by 10 folds) one has to increase the MCMC sample size by 100 folds.

3.1 Analysis A00: Parameter estimation under the multispecies coalescent

This analysis (with `speciesdelimitation = 0`, `speciestree = 0`) generates the posterior distribution of species divergence times (τs) and population sizes (θs) under the MSC model when the species phylogeny is fixed. As noted above, parameters θs and τs are the products of time and mutation rate. The sequence data provide information about distances only, so that time and rate are confounded. If external information is available concerning the mutation rate or if fossil information can be used to calibrate the ages of nodes on the species tree, the estimates of θs and τs can be converted into estimates of absolute species divergence times and absolute population sizes (see, e.g., Burgess and Yang, 2008). Such an analysis under the MSC accommodates ancestral polymorphism and the coalescent waiting times in the ancestral species and may be advantageous over traditional molecular clock dating (Angelis and dos Reis, 2015). In the case where a fossil calibration is available for the root of the species tree (in the form of sharp minimum and soft maximum bounds, say), Angelis and dos Reis (2015) discuss a strategy of sampling from the prior calibration distribution to generate posterior estimates of species divergence times and population sizes as well as the mutation rate.

Run the program as follows.

```
cd frogs\r1
..\..\bpp ..\A00.bpp.ctl
```

The control file `A00.bpp.ctl` is shown in Fig. 3. Here clades C and K are treated as distinct species, and the fixed species phylogeny is ((H, L), (C, K)). This is the most favoured model in the analysis below although we note that the phylogeny is highly uncertain. The parameters in the model are defined in Fig. 2A. We assign a gamma prior G(2, 1000) for all θ parameters, and $\tau_0 \sim$ G(2, 2000) for the age of the root, while the other divergence time parameters are assigned the uniform Dirichlet prior (Yang and Rannala, 2010: equation 2). The prior $\theta \sim$ G(2, 1000) has the mean 2/1000 = 0.002, which means 2 differences per kilobase between two sequences sampled at random from the population. Similarly the prior $\tau_{KCLH} \sim$ G(2, 2000) has the mean 0.001, which means that the sequences at the root and a tip of the tree have one difference per kilobase. The uniform Dirichlet prior for the other τs then means that given τ_{KCLH}, the ages τ_{KC} and τ_{LH} are uniform in the interval $(0, \tau_{KCLH})$.

The gamma priors. In BPP, gamma priors are used on θs and τ. A gamma distribution is specified as G(α, β), with shape parameter α and rate parameter β, and with

mean α/β and variance α/β^2. The shape parameter α specifies how informative the prior is, with α = 1, 1.5 or 2 representing diffuse priors, while values like 10, 20, and 100 represent informative priors. There are no default priors for parameters θ and τ in BPP, and the priors should be chosen to suit the dataset being analyzed. One way of specifying the gamma prior is to choose α depending on how much information one has about the parameters and then to choose β so that the prior means are in the right neighbourhood. Here we use α = 2 to have a diffuse prior. Rough estimates of the parameters can be generated by a preliminary run. For example, if we use the priors $\theta \sim$ G(2, 500) and $\tau_{\text{KCLH}} \sim$ G(2, 500), both with mean 0.004, most of the posterior means of θs are <0.004, and that of τ_{KCLH} is around 0.001, suggesting that those prior means are too large. In contrast, if we use the priors $\theta \sim$ G(2, 2000) and $\tau_{\text{KCLH}} \sim$ G(2, 2000), both with mean 0.001, the posterior means of all θs are >0.001, suggesting that the prior mean for θs is too small, and the posterior mean of τ_{KCLH} is around 0.0007, suggesting that the prior mean for τ_{KCLH} is slightly too large.

Note that the prior is supposed to represent our information about the parameters before the analysis of the data. It is thus incorrect to fit the gamma distribution to the posterior sample and use the resulting gamma distribution as the prior. Here we have chosen to use a diffuse prior (with shape parameter α = 1 or 2), and the preliminary runs are used to ensure that the prior means are reasonable for the data.

In specifying the gamma priors, it may also be useful to plot the gamma density and calculate the 95% prior interval. The following R commands plot the gamma density G(2, 2000) and calculates the 95% prior interval to be (0.00012, 0.00279).

```
a=2; b=2000;
curve(dgamma(x, a, b), from=0, to=0.01)
qgamma(c(0.025, 0.975), a, b)
```

Edit the control file to use `usedata = 0`. This sets the sequence likelihood to 1 whatever the parameter values, so that the MCMC will generate a sample from the prior. Confirm that the means calculated by BPP are 0.002 for all θs, 0.001 for τ_{KCLH}, and 0.0005 for both τ_{KC} and τ_{LH}, as specified in the priors.

Then edit the control file to use `usedata = 1` and rerun the program to sample from the posterior. The MCMC sample file (`mcmc.txt`) from this analysis can be read into R or Tracer to plot the estimated posterior densities. (Note, however, that the sample files from the other analyses A01, A10 and A11 are not readable in TRACER.) Here we use the summary provided by BPP. The output should be self-explanatory, and is summarized in Fig. 2B. The posterior means of the divergence times (τs) are used to draw the branches of the tree, which also shows the 95% highest probability density (HPD) intervals as node bars. The posterior means of the population size parameters (θs) are shown along the branches, which range from 0.0016 to 0.0042. Overall, the parameter estimates have large intervals, indicating that the information content in the five nuclear loci is quite low.

The reader may wish to examine the sensitivity of the posterior estimates to the priors by changing the parameters in the prior, for example, by using $\theta \sim$ G(2, 100) and $\tau \sim$ G(2, 200), with the prior means to be ten times as large. The prior means are expected to have more impact than the shape parameters.

Note that the MSC model assumes that the samples are taken at random from each population or species. Thus all sequences generated should be included in the analysis even if some of them are identical. It is incorrect to use the distinct haplotypes only, as removal of the identical sequences leads to overestimates of the parameters (θs and τs).

3.2 Analysis A01: Species tree estimation

This analysis (with `speciesdelimitation = 0`, `speciestree = 1`) infers the species tree, assuming that the assignment and species delimitation are fixed. Based on the MSC, the analysis accounts for polymorphism in the ancestral species and the resultant gene tree-species tree conflicts. It also accommodates the uncertainties in the gene trees due to limited phylogenetic information at each locus. BPP uses the nearest neighbor interchange (NNI) or subtree pruning and regrafting (SPR) algorithms to change the species tree topology in the MCMC (Yang and Rannala, 2014). We run the program as follows.

```
cd frogs\r1
..\..\bpp ..\A01.bpp.ctl
```

Bayesian species tree estimation is essentially a model selection analysis since different species phylogenies are different statistical models. Thus we have to specify prior probabilities for the compared models (species trees). Two priors are implemented in BPP for this analysis. Prior 0 (`speciesmodelprior = 0`)

assigns equal probabilities for the labeled histories (which are rooted trees with the internal nodes ordered by age), while Prior 1 (`speciesmodelprior = 1`) assigns equal probabilities for the rooted trees (Yang and Rannala, 2014). For instance, there are 15 rooted trees in the case of four species (A, B, C, and D), with 12 unbalanced and 3 balanced trees. Each unbalanced tree, e.g., (((A, B), C), D), is compatible with only one labeled history as there is only one ordering of the internal nodes. Each balanced tree, e.g., ((A, B), (C, D)), is compatible with two labeled histories, depending on whether the ancestor of A and B is older or younger than the ancestor of C and D. Prior 0 assigns the probability 1/18 to each of the unbalanced trees and 2/18 to each of the balanced trees. Prior 1 assigns the probability 1/15 to each of the 15 rooted trees. We use Prior 1, which is the default. Within each species tree model we assign the gamma priors $\theta \sim$ G(2, 1000) for all θs and $\tau \sim$ G(2, 2000) for the root age.

The program collects the species trees (as well as parameters θs and τs) into the sample file `mcmc.txt`. The BPP summary of the MCMC sample is shown in figure 4, which shows the top four species trees and their posterior probabilities. Those trees have a total posterior probability of 0.5 and thus constitute the 50% credibility set. The 95% credibility set includes 13 trees, while the 99% credibility set includes all the 15 possible trees. The majority-rule consensus tree is the star phylogeny. Overall there is very limited phylogenetic information in the five nuclear loci, due to the low levels of sequence divergence.

It has been noted that a large prior mean for θ makes the different species tree look similar and thus reduces the posterior probabilities for trees. The reader may wish to explore the impact of the prior for θs and τs on the posterior probabilities. Change the prior means so that they are a few times too large or a few times too small. For example, using $\theta \sim$ G(2, 100) and $\theta \sim$ G(2, 10,000) will allow the prior mean to vary over two orders of magnitude, which should be more than enough. Similarly you can use $\tau \sim$ G(2, 200) and $\tau \sim$ G(2, 20,000).

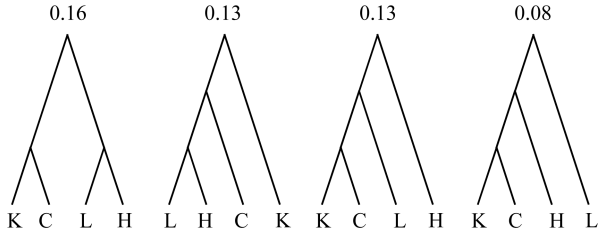

Fig. 4 Analysis 01

The top four species trees and their posterior probabilities, with a total probability of 0.5. The priors are $\theta \sim$ G(2, 1000) for all populations and $\tau \sim$ G(2, 2000) for the root age.

3.3 Analysis A10: species delimitation on a guide tree

In this analysis (with `speciesdelimitation = 1, speciestree = 0`), a reversible-jump MCMC (rjMCMC) algorithm is used to move between different species-delimitation models that are compatible with a fixed guide tree (Yang and Rannala, 2010; Rannala and Yang, 2013). We run the program as follows.

```
cd frogs\r1
..\..\bpp ..\A10.bpp.ctl
```

The control file uses the guide tree ((K, C), (L, H)), shown in Figure 2A. The rjMCMC algorithm will attempt to collapse some of the internal nodes on the guide tree but will not change the guide tree. A collapsed node means that the descendent populations of the node all belong to the same species. Thus five models of species delimitation and species phylogeny will be explored by the algorithm, all specified by the guide tree. These are KCLH (coded 000, 1 species), KC-LH (coded 100, 2 species), KC-L-H (coded 101, 3 species), K-C-LH (coded 110, 3 species), and K-C-L-H (coded 111, 4 species). The models and their prior probabilities are listed on the screen at the start of the run, as follows.

```
Number of species-delimitation models =  5
       delimitation model   1: 000  prior  0.20
       delimitation model   2: 100  prior  0.20
       delimitation model   3: 101  prior  0.20
       delimitation model   4: 110  prior  0.20
       delimitation model   5: 111  prior  0.20

[Note: Ancestral nodes in order:  5 KCLH  6 KC  7 LH]
```

Here the model is coded using three 0-1 flags, which indicate whether the three ancestral nodes (5, 6, and 7 in fig. 2a) are present (flag 1) or absent (flag 0). For example the fourth model (K-C-LH), coded 110, means that nodes 5 and 6 are present and their daughter nodes represent distinct species while node 7 is collapsed and its daughter nodes L and K are one species. When there are three or more delimited species in the model, the species tree is also fixed by the guide tree. For example, the 4th model K-C-LH (coded 110) has 3 species, and the species tree is (KC, (L, H)), as given by the guide tree.

Two alternative rjMCMC algorithms are implemented in BPP, specified as follows.

```
speciesdelimitation = 1 0 2. * speciesdeli-
mitation algorithm0 and finetune(e)
```

```
speciesdelimitation = 1 1 2 1 * speciesdelimitation algorithm1 finetune (a m)
```

The first is Algorithm 0, with $\varepsilon = 2$ in equations 3 and 4 of Yang and Rannala (2010). Reasonable values for ε are 1, 2, 5, etc. The second is Algorithm 1, with $\alpha = 2$ and $m = 1$ in equations 6 and 7 of Yang and Rannala (2010). Reasonable values are $\alpha = 1$, 1.5, 2, etc. and $m = 0.5$, 1, 2, etc. When the chain mixes well, the results should be the same between multiple runs using the two algorithms.

We use the default prior for the different species tree models (`speciesmodelprior = 1`), which assigns equal probabilities for the rooted trees. We assign the same priors on θs and τs as before: $\theta \sim$ G(2, 1000) for θs and $\tau \sim$ G(2, 2000) for the root age in every species phylogeny.

The program collects the sampled delimitation model and the parameters in the model (θs and τs) in the sample file `mcmc.txt`. This is summarized by the program. The model of four species has posterior probability 0.87, while the three-species model that groups L and H into one species has posterior probability 0.13. The other models have negligible probabilities.

In theory the transmodel MCMC generates both the posterior probabilities of the models and the posterior distribution of the parameters (τs and θs) within each model. The latter can be estimated by using only those samples in which the chain is in that particular model. This within-model parameter posterior can also be generated by running the simple MCMC (analysis A00) with the species delimitation and species tree fixed at that particular model (which can be achieved by editing the Imap and control files). We recommend this latter approach. This applies to all the transmodel inferences discussed in this paper.

It has been noted that use of a large prior mean for θ makes BPP produce unresolved trees and/or lump populations into the same species (e.g., Leaché and Fujita, 2010; Zhang et al., 2011; Pelletier et al., 2015). We leave it to the reader to explore the sensitivity of the posterior model probabilities to the prior specification. The reader may also wish to change the guide tree to examine its impact. Note that it is not sensible to average the posterior model probabilities over different priors or over different runs (if some runs have failed to converge), as in Pelletier et al. (2015).

3.4 Analysis A11: Joint species delimitation and species-tree estimation

In this analysis (with `speciesdelimitation = 1`, `speciestree = 1`), the algorithm explores different species delimitation models and different species phylogenies. The assignment of individuals to populations is nevertheless fixed: the program attempts to merge different populations into one species but never tries to split one population into multiple species. The nearest neighbor interchange (NNI) or subtree pruning and regrafting (SPR) algorithms are used to change the species tree topology, while rjMCMC is used to split one species into two or to join two populations into one species (Yang and Rannala, 2014). Run the program as follows.

```
cd frogs\r1
..\..\bpp ..\A11.bpp.ctl
```

For analysis A11, BPP 3.1 provides four priors which assign probabilities to models. They are referred to as Priors 0, 1, 2 and 3. Prior 0 assigns equal probabilities for the labeled histories (rooted trees with the internal nodes ordered by age), while Prior 1 assigns equal probabilities to the rooted species trees. Those two priors were implemented by Yang and Rannala (2014) and have been mentioned above. Priors 2 and 3 assign equal probabilities for the numbers of species ($1/s$ each for 1, 2, ..., and s delimited species given s populations) and then divide up the probability for any specific number of species among the compatible models (of species delimitation and species phylogeny) either in proportion to the number of compatible labeled histories (Prior 2) or uniformly (Prior 3). Priors 2 and 3 are mentioned by Yang and Rannala (2014) and but not implemented until this version. A detailed description of Priors 2 and 3 for the cases of four or five populations is given in Table 2. Prior 3 may be suitable when there is a large number of populations. One such scenario is when each individual (specimen) is assigned into its own "population", so that BPP will explore different models of assignment, species delimitation and species tree estimation (Olave et al., 2014). Here in this tutorial, we use Prior 1.

Within each model we assign the priors $\theta \sim$ G(2, 1000) for all θs and $\tau \sim$ G(2, 2000) for the root age in each species tree when there are two or more species in the model. The species tree in the control file is used as the starting guide tree. When the algorithm converges, use of different starting guide trees should lead to the same results.

From this run, the posterior probability for four species (K, C, L, H) is 0.95, while that for three species is 0.05 (0.03 for joining L and H into one species and 0.01 for joining K and H). There is evidence in the dataset

Table 2 Prior probabilities for models specified by Priors 2 and 3 for the cases of four and five populations (for Analysis A11)

Prior 2	Prior 3
(**a**) The case of four populations (A, B, C, D) (41 models in total)	
P_1 = 0.25: 1 delimitation ABCD	
P_2 = 0.25: 4 delimitations of form A-BCD (each 0.05), 3 delimitations of form AB-CD (each 0.05/3 = 0.0167),	
P_3 = 0.25: 6 delimitations of form A-B-CD, 3 trees for each (18 models, each 0.25/18 = 0.0139)	
P_4 = 0.25: 1 delimitation A-B-C-D, 15 trees (12 unbalanced trees, each 0.25/18 = 0.0139, and 3 balanced trees, each 0.25/9 = 0.0278)	P_4 = 0.25: 1 delimitation A-B-C-D, 15 trees (each 0.25/15 = 0.0167)
(**b**) The case of five populations (A, B, C, D, E) (346 models in total)	
P_1 = 0.2: 1 delimitation ABCDE	
P_2 = 0.2: 5 delimitations of form A-BCDE (each 0.2/105×15 = 0.0286), 10 delimitations of form AB-CDE (each 0.2/105 × 3 = 0.0056),	
P_3 = 0.2: 10 delimitations of form A-B-CDE, 3 trees for each (30 models, each 0.2/135 × 3 = 0.00444); 15 delimitations of form A-BC-DE, 3 trees for each (45 modes, each 0.2/135 = 0.00148)	
P_4 = 0.2: 10 delimitations of form A-B-C-DE, each 15 trees (10 × 12 trees, each 0.2/180 = 0.00111 and 10 × 3 trees each 0.00222)	P_4 = 0.2: 10 delimitations of form A-B-C-DE, each 15 trees (150 models, each 0.2/150 = 0.00133)
P_5 = 0.2: 1 delimitation A-B-C-D-E, 105 trees: 60 trees of form (((A, B), C), D), E) (each 0.2/180 = 0.00111); 30 trees of form ((A,B)((C,D)E)) (each 0.00333); 15 trees of form (((A, B)(C, D))E) (each 0.00222)	P_5 = 0.2: 1 delimitation A-B-C-D-E, 105 trees (0.2/105= 0.00190)

Note: Species delimitation is indicated using dashes, and species tree using the parenthesis notation. Both Priors 2 and 3 assign equal prior probabilities to the different numbers of species (1/4 each for four populations and 1/5 each for five populations). For example, in the case of five populations, Prior 3 partitions P_4 = 0.2 for four species among the 150 compatible models uniformly, with each receiving probability 0.2/150 = 0.00133. Prior 2 partitions P_4 = 0.2 in proportion to the number of labeled histories, so that 120 unbalanced tree models are assigned the probability 0.2/180 each while 30 balanced tree models receive 0.2/90 each. Priors 2 and 3 assign the same prior probabilities for models of 1, 2, and 3 species.

for distinct species status for the two *R. chensinensis* populations (C and L), although the evidence is not very strong. It has been suggested that different populations be declared distinct species only if the posterior probability exceeds a threshold such as 95% or 99% (Rannala and Yang, 2013). The phylogenetic relationships among the delimited species are highly uncertain, with the 95% credibility set including as many as 15 models. The data seem to contain far more information about species delimitation than about species phylogeny. This is also the pattern found in the analyses of other datasets (Yang and Rannala, 2014; Caviedes Solis et al., 2015). Thus the unguided delimitation (A11) should be preferred over species delimitation using a fixed guide tree (A10).

Mitochondrial introgression. In all analyses described above, the three *R. chensinensis* samples involved in mitochondrial introgression are assigned to the C clade. Here we conduct a joint species-delimitation and species tree estimation by assigning those samples to a population of their own (m). The modified control file is named `A11.bpp.introgression.ctl`, which defines five populations instead of four, and uses the Imap file `frogs.Imap.introgression.txt`. Run BPP as follows.

```
..\..\bpp ..\A11.bpp.introgression.ctl
```

The posterior probability for five delimited species (K, m, C, L, H) is 0.71, and the probability for 4 species is 0.28 (0.24 for merging m and C into one species and 0.02 for merging L and H into one species), and the probability for 3 species is 0.01. The maximum *a posteriori* probability (MAP) model is the five-species model ((L, H), (K, (m, C))), but the posterior probability is very low, at 8%. Given the species delimitation, the species phylogeny is highly uncertain.

Table 3 summarizes results obtained from using different priors on θs and τs. The prior mean for θs has considerable influence on the Bayesian model selection, with larger prior means favouring fewer species. In comparison, the prior on τ has much less impact. For example, with the priors $\theta \sim$ G(2, 100) and $\tau \sim$ G(2, 200), in which case the prior means are 10 times too large, the posterior probabilities for 5, 4, and 3 species become 0.52, 0.41, and 0.07, compared with 0.71, 0.28, and 0.01 discussed above for the prior $\theta \sim$ G(2, 1000) and $\tau \sim$ G(2, 2000). The MAP model is ((L, H), (K, (m, C))), with the posterior probability 8%. Overall, the prior on parameters influences the posterior probabilities of the models but does not change the ranking of the models.

Table 3 Posterior probabilities for the number of delimited species using different priors for model parameters (Analysis A11)

Prior	Posterior probability for the number of delimited species
(a) Using four populations (K, C, L, H)	
$\theta \sim$ G(2, 1000), $\tau \sim$ G(2, 2000)	$P_4 = 0.95$, $P_3 = 0.05$
$\theta \sim$ G(2, 100), $\tau \sim$ G(2, 200)	$P_4 = 0.83$, $P_3 = 0.17$
$\theta \sim$ G(2, 100), $\tau \sim$ G(2, 2000)	$P_4 = 0.81$, $P_3 = 0.19$
$\theta \sim$ G(2, 1000), $\tau \sim$ G(2, 200)	$P_4 = 0.96$, $P_3 = 0.04$
(b) Using five populations (K, m, C, L, H)	
$\theta \sim$ G(2, 1000), $\tau \sim$ G(2, 2000)	$P_5 = 0.71$, $P_4 = 0.28$, $P_3 = 0.01$
$\theta \sim$ G(2, 100), $\tau \sim$ G(2, 200)	$P_5 = 0.52$, $P_4 = 0.41$, $P_3 = 0.07$
$\theta \sim$ G(2, 100), $\tau \sim$ G(2, 2000)	$P_5 = 0.53$, $P_4 = 0.40$, $P_3 = 0.07$
$\theta \sim$ G(2, 1000), $\tau \sim$ G(2, 200)	$P_5 = 0.69$, $P_4 = 0.29$, $P_3 = 0.02$

Note: Three individuals in the data appear to be involved in mitochondrial introgression, and are assigned to population C in the 4-population analysis (a) and to a population of their own (m) in the 5-population analysis (b). Prior 1, which assumes uniform probabilities for the rooted trees, is assumed.

4 Discussion

4.1 Assumptions, strengths and weaknesses of the bpp analysis

Here we provide a brief discussion of the assumptions made in all BPP analyses described in this paper and their possible effects, as well as the strengths and weaknesses of the BPP analysis compared with some of the alternatives. The analysis in BPP makes the following standard assumptions: (i) no recombination among sites within a locus and free recombination between loci; (ii) neutral clock-like evolution at each locus and JC69 mutation model (Jukes and Cantor, 1969); and (iii) no migration (gene flow) between species.

The assumptions concerning recombination suggest that suitable sequence data for the program should be short segments of the genome that are loosely linked, such that recombination among sites within each segment is negligible while the segments are far apart so that they are nearly freely recombining. For organisms with very high recombination rates, the assumption of no recombination within a locus may be seriously violated. Lanier and Knowles (2012) has examined the impact of recombination on species tree inference using *BEAST (Heled and Drummond, 2010) and STEM (Kubatko et al., 2009) and found that it had only minimal impact. The effect of recombination on species delimitation may be similar.

The assumption of neutral evolution at the loci may not be as important as is often suggested in the literature. Protein-coding gene sequences should be useable in a BPP analysis if the proteins are performing similar functions in the different species and under similar selective constraints. Such purifying selection has the effect of reducing the neutral mutation rate, although it may be necessary to accommodate the variation in mutation rate among loci in the analysis. In contrast, species-specific directional selection may distort the shape of the genealogical tree and have more serious impact on species delimitation and species tree estimation. Analysis of such loci using BPP should proceed with caution. It may be useful to examine the posterior distribution of the gene trees since directional or positive selection may be expected to lead to unusual gene trees. The major role of the mutation model in the analysis is to correct for multiple hits at the same site. If the species and populations are closely related and the sequences are highly similar, the JC69 model should be adequate. This may be the case for species delimitation (analyses A10 and A11). However for species tree estimation when the species are divergent (analysis A01), as in the case of phylogeny estimation for placental mammals (Gatesy and Springer, 2013), both the molecular clock and the JC69 model may be seriously violated. Use of BPP in such cases is not advisable. Relaxed-clock models and sophisticated nucleotide-substitution models are yet to be implemented in BPP.

The models implemented in BPP ignore possible migration between populations or species. This means that one cannot use BPP to estimate the migration rates between species. Migration is also expected to influence species delimitation, and should homogeneize the species, causing BPP to lump distinct species into one species. The simulation conducted by Zhang et al. (2011) suggests that BPP behaves sensibly in presence of migration. When migration rate is low, with < 0.1 migrants per generation, say, migration is found to have little effect on species delimitation by BPP. However, when the migration rate is very high, with as many as 10 migrants per generation, BPP tends to infer one species. Thus BPP appears to behave like a pragmatic taxonomist. The impact of migration on species tree estimation has been evaluated by Leaché et al. (2014). The situation is complex. For example, gene flow may either hinder or improve species tree estimation, depending on which species are exchanging migrants.

Given the assumptions discussed above, BPP is a full likelihood-based implementation of the MSC model. The gene tree topologies and branch lengths (coalescent

times) have independent probability distributions among loci specified by the species tree and parameters (θs and τs) (Rannala and Yang, 2003). Given the gene tree topology and branch lengths at each locus, the likelihood (or the probability of the sequence alignment at the locus) is calculated using Felsenstein's (1981) pruning algorithm under the JC9 mutation model. See Yang (2014: Chapters 6 and 9) for an introduction to the MCMC computational algorithms. The implementation in BPP is thus in contrast to the heuristic or short-cut coalescent methods, which typically involve using phylogenetic methods to estimate the gene trees and then using the estimated gene trees to infer the species tree, without accommodating properly the uncertainties in the estimated gene trees. Some of those heuristic methods use the gene tree topologies and ignore information in the branch lengths, leading to dramatic loss of information (Gatesy and Springer, 2013).

The advantages of the BPP analysis have been discussed by a number of authors (Fujita and Leaché, 2011; Fujita et al., 2012; Yang, 2014: Chapter 9; Rannala, 2015). The program accounts for ancestral polymorphism and incomplete lineage sorting. It makes a full use of the information in the sequence data, and accommodates the uncertainties in the topologies and branch lengths in the gene trees at the individual loci. The latter feature may be important for species delimitation and species tree estimation when the involved species are closely related and the sequences at any locus are highly similar and thus contain little phylogenetic information.

4.2 What to report in your study

Almost all studies using MRBAYES (Ronquist et al., 2012) or BEAST (Drummond and Rambaut, 2007) report the number of iterations, but very few clearly specify the priors used in the analysis. However, the number of iterations is neither necessary nor sufficient to guarantee the success of an MCMC run. Nor is it required for reproducing the analysis. Different MCMC programs use very different algorithms so that an iteration in one program is not comparable with an iteration in another. For example, in MRBAYES and BEAST, one iteration may be equivalent to sampling one parameter in the model for updating, while one iteration in BPP may be equivalent to updating all parameters in the model one by one. In that case one iteration in BPP may be worth 10^3 or 10^4 iterations in MRBAYES or BEAST. In contrast, knowledge of the prior specification is necessary for reproducing a Bayesian analysis. Thus we encourage the reporting of the specification of the prior.

4.3 Mixing problems with the MCMC algorithms in bpp

As mentioned above, analyses A01, A10, and A11 are transmodel inferences, and in those analyses, BPP in effect conducts a standard Bayesian model selection. Each of those models is an instance of the MSC model, but the number and nature of the species and the species phylogeny may differ among those models. For example, in the species-delimitation analysis (A10) of the tutorial, both models KC-L-H (3 species) and K-C-LH (3 species) have three species but they are not the same three species.

For the example dataset of the frogs, the BPP algorithms appear to have worked well in all four analyses. Nevertheless, it has been noted that the transmodel inferences (A01, A10, and A11) may suffer from mixing problems in some datasets. The mixing problem here concerns the efficiency but not the correctness of the algorithm. A correct MCMC algorithm should visit the different models in proportion to their posterior probabilities. However, an efficient algorithm may jump between models frequently while an inefficient (lazy) algorithm may stay in one model for a long time before it jumps, and then stays in the new model for a long time before it jumps. Both algorithms are correct in the sense that in the long run they both visit the models in proportion to the posterior probabilities. However, the lazy algorithm may be very inefficient as it takes an extremely long chain to generate reliable results. The main symptom for poor mixing of the transmodel algorithm is that the chain gets stuck in one model (or a subset of models), and multiple runs (each over a finite number of iterations) produce different results. Mixing problems tend to be worse and occur more frequently for larger datasets but can occur even for small datasets. It has been noted that multiple runs using different starting species trees or species delimitation models are effective in exposing mixing problems, and that consistency of results among multiple runs is a good indication for the success of the run. It is thus important to conduct multiple runs for the transmodel inferences.

Acknowledgments I thank Dr Weiwei Zhou for providing the brown frog sequence data and Drs Hui Zhao and Jing Che for the frog photos. Dr Chi Zhang conducted the initial analysis of the frog data. I am grateful to Nassima Bouzid, Itzue Caviedes-Solis, Adam Leaché, Cheng-Min Shi, Chi Zhang, and Tianqi Zhu for many constructive comments on early versions of this paper. This study is supported by a grant from the Biotechnological and Biological Sciences Research Council (BBSRC) to Z.Y.

References

Angelis K, dos Reis M, 2015. The impact of ancestral population size and incomplete lineage sorting on Bayesian estimation of species divergence times. Curr. Zool. 61: 874–885.

Burgess R, Yang Z, 2008. Estimation of hominoid ancestral population sizes under Bayesian coalescent models incorporating mutation rate variation and sequencing errors. Mol. Biol. Evol. 25: 1979–1994.

Carstens BC, Pelletier TA, Reid NM, Satler JD, 2013. How to fail at species delimitation? Mol. Ecol. 22: 4369–4383.

Caviedes Solis I, Bouzid N, Banbury B, Leaché AD, 2015. Uprooting phylogenetic uncertainty in coalescent species delimitation: A meta-analysis of empirical studies. Curr. Zool. 61: 866–873.

Drummond AJ, Rambaut A, 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7: 214.

Edwards SV, 2009. Is a new and general theory of molecular systematics emerging? Evolution 63: 1–19.

Edwards SV, Liu L, Pearl DK, 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. U.S.A. 104: 5936–5941.

Felsenstein J, 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17: 368–376.

Fujita MK, Leaché AD, 2011. A coalescent perspective on delimiting and naming species: A reply to Bauer et al. Proc. R. Soc. Lond. B. Biol. Sci. 278: 493–495.

Fujita MK, Leaché AD, Burbrink FT, McGuire JA, Moritz C, 2012. Coalescent-based species delimitation in an integrative taxonomy. Trends Ecol. Evol. 27: 480–488.

Gatesy J, Springer MS, 2013. Concatenation versus coalescence versus "concatalescence". Proc. Natl. Acad. Sci. U.S.A. 110: E1179.

Green, P. J. 2003. Trans-dimensional Markov chain Monte Carlo. In: Green PJ, Hjort NL, Richardson S ed. Highly Structured Stochastic Systems. Oxford: Oxford University Press, 179–196

Heled J, Drummond AJ, 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27: 570–580.

Hey J, 2010. Isolation with migration models for more than two populations. Mol. Biol. Evol. 27: 905–920.

Hobolth A, Christensen OF, Mailund T, Schierup MH, 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3: e7.

Jukes TH, Cantor CR, 1969. Evolution of protein molecules. In: Munro HN ed. Mammalian Protein Metabolism. New York: Academic Press, 21–123.

Kubatko LS, Carstens BC, Knowles LL, 2009. STEM: Species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25: 971–973.

Lanier HC, Knowles LL, 2012. Is recombination a problem for species-tree analyses? Syst. Biol. 61: 691–701.

Leaché AD, Fujita MK, 2010. Bayesian species delimitation in West African forest geckos *Hemidactylus fasciatus*. Proc. R. Soc. Lond. B. Biol. Sci. 277: 3071–3077.

Leaché AD, Harris RB, Rannala B, Yang Z, 2014. The influence of gene flow on Bayesian species tree estimation: A simulation study. Syst. Biol. 63: 17–30.

Liu L, Pearl DK, 2007. Species trees from gene trees: Reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56: 504–514.

Liu L, Xi Z, Wu S, Davis C, Edwards SV, 2015. Estimating phylogenetic trees from genome-scale data. Ann. NY. Acad. Sci. doi: 10.1111/nyas.12747

Olave M, Sola E, Knowles LL, 2014. Upstream analyses create problems with DNA-based species delimitation. Syst. Biol. 63: 263–271.

Pelletier TA, Crisafulli C, Wagner S, Zellmer AJ, Carstens BC, 2015. Historical species distribution models predict species limits in Western Plethodon salamanders. Syst. Biol.: in press.

Rannala B, 2015. Species delimition and species tree estimation using multilocus sequence data. Curr. Zool. 61: 846–853.

Rannala B, Yang Z, 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.

Rannala B, Yang Z, 2013. Improved reversible jump algorithms for Bayesian species delimitation. Genetics 194: 245–253.

Robert CP, Casella G, 2004. Monte Carlo Statistical Methods, 2nd edn. New York: Springer-Verlag.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A et al., 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61: 539–542.

Takahata N, Satta Y, Klein J, 1995. Divergence time and population size in the lineage leading to modern humans. Theor. Popul. Biol. 48: 198–221.

Yang Z, 2002. Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. Genetics 162: 1811–1823.

Yang Z, 2014. Molecular Evolution: A Statistical Approach. Oxford: Oxford University Press.

Yang Z, Rannala B, 2010. Bayesian species delimitation using multilocus sequence data. Proc. Natl. Acad. Sci. U.S.A. 107: 9264–9269.

Yang Z, Rannala B, 2014. Unguided species delimitation using DNA sequence data from multiple loci. Mol. Bio.l Evol. 31: 3125–3135.

Zhang C, D-X Zhang, Zhu T, Yang Z, 2011. Evaluation of a Bayesian coalescent method of species delimitation. Syst. Biol. 60: 747–761.

Zhou WW, Wen Y, Fu J, Xu YB, Jin JQ et al., 2012. Speciation in the *Rana chensinensis* species complex and its relationship to the uplift of the Qinghai-Tibetan Plateau. Mol. Ecol. 21: 960–973.