BPEC: An R Package for Bayesian Phylogeographic and Ecological Clustering

Ioanna Manolopoulou
University College London

Axel Hille
Institute of Applied Statistics
Dr Jörg Schnitker

Brent Emerson
Instituto de
Productos Naturales
y Agrobiologia, and
University of East Anglia

Abstract

BPEC is an R package for Bayesian Phylogeographic and Ecological Clustering which allows geographical, environmental and phenotypic measurements to be combined with DNA sequences in order to reveal geographic structuring of DNA sequence clusters consistent with migration events. DNA sequences are modelled using a collapsed version of a simplified coalescent model projected onto haplotype trees, which subsequently give rise to constrained clusterings as migrations occur. Within each cluster, a multivariate Gaussian distribution of the covariates (geographical, environmental, phenotypic) is used. Inference follows tailored Reversible Jump Markov chain Monte Carlo sampling so that the number of clusters (i.e., migrations) does not need to be pre-specified. A number of output plots and visualizations are provided which reflect the posterior distribution of the parameters of interest. **BPEC** also includes functions that create output files which can be loaded into Google Earth. The package commands are illustrated through an example dataset of the polytypic Near Eastern brown frog *Rana macrocnemis* analysed using **BPEC**.

Keywords: statistical phylogeography, biogeography, population genetics, Bayesian computation, R.

1. Introduction

Phylogeography can be considered the nexus between classical population genetics, phylogenetics and historical biogeography, with much conceptual and analytical overlap with all three, but particularly with population genetics. Phylogeography was born from the integration of population genetics and phylogenetics to work at the micro- macroevolutionary interface (Hickerson *et al.* 2010), being an evolved discipline that seeks to integrate the genealogical relationships among DNA lineages (sequences) with their geographic distributions to infer historical events that have shaped the contemporary distributions of species and their genetic variation. However, while population genetics, phylogenetics and historical biogeography have witnessed a growth of analytical approaches in recent years, there has been a relative dearth of analytical approaches within the field of phylogeography, with several reviews summarising these (e.g. Knowles 2009; Bloomquist *et al.* 2010; Hickerson *et al.* 2010). To place our work into a broader context, we provide a brief summary of the state of the art within the field of phylogeography, but the aforementioned reviews should be referred to for

more detail.

Historical biogeography seeks to understand the processes that have shaped the evolution of geographic differences among related species (i.e., interspecific process), and may involve timescales that extend back tens of million of years or more. In contrast, phylogeography concerns both the quantification of the geographic structuring of genetic variation within species, and understanding the process that has shaped said structure (i.e. intraspecific process). Thus phylogeographic analyses typically involve time-scales that don't extend back more than a few million years. A challenge for phylogeographic analysis is to simultaneously account for evolutionary processes over spatial and temporal dimensions, and perhaps for this reason the phylogeographer's toolkit is a mixed bag of approaches encompassing various objectives within this framework. Some population genetic methods find relevance in phylogeography, precisely because they do not use geographical information explicitly, but rely on population genetics modelling to infer the geography of structure. For example, **STRUCTURE** (Pritchard *et al.* 2000) infers population structure purely from genotype data through a Latent Dirichlet Allocation model. Population subdivisions are assessed on the basis of multi-locus allele frequencies which are directly learnt from data. More recently, Jombart *et al.* (2010) developed **DAPC**, a principal-components alternative to **STRUCTURE** which can computationally efficiently deal with large amounts of data. In these approaches one describes genetic groupings in the absence of spatial information, onto which phylogeographic inferences can then be conditioned. Fully model-based extensions of spatially-explicit inferences of population structure such as **GENELAND** (Guillot *et al.* 2005) and Cheng *et al.* (2013) assume that the spatial domain occupied by the inferred clusters can be approximated by a small number of polygons based on Voronoi tessellations. Drawing inferences about these cluster domains (and thus about cluster membership) amounts to inferring the location and cluster memberships of the polygons. Finally, recent approaches such as Jay *et al.* (2015) introduced spatially-dependent cluster membership probabilities through a regression model. These approaches use multilocus genotype data for the inference of spatial genetic structure, and therefore the absence of a coalescent framework limits inferences across the temporal dimension.

Methods that use the evolutionary relationships among alleles for phylogeographic analysis open the door for jointly investigating the spatial and temporal dimensions of genetic relatedness among individuals. Early phylogeography relied upon qualitative assessments of the geographic relationships within a gene genealogy, together with estimated dates of gene tree branching events. In this approach demography was directly inferred from the phylogenetic relationships of alleles, with limited importance given to the potentially confounding effects of coalescent stochasticity (Hickerson *et al.* 2010). Such stochasticity could give rise to similarly probable alternative demographic explanations for a given data set. To address this, simulation-based statistical methods based on coalescent models for parameter estimation have emerged giving rise to statistical phylogeography (Knowles and Maddison 2002; Knowles 2009) allowing for testing among competing demographic models.

With regard to the joint analysis of the genealogical and spatial relationships of DNA sequences, we are only aware of two implementations to date. (Lemey *et al.* 2009) developed a fully model-based Bayesian phylogeographic inference framework, assuming a diffusion model for the geographical migration of nodes on a phylogenetic tree, so that evolution and migration events occur in a continuous-time framework. More recently, Guindon *et al.* (2016) modelled spatial distribution as a gradual dispersal across a continuous landscape.

Here we present a novel R (R Core Team 2019) package available from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=BPEC which automates Bayesian Phylogeographic and Ecological Clustering (**BPEC**) analysis (Manolopoulou *et al.* 2011; Manolopoulou and Emerson 2012). **BPEC** is a model-based approach which assumes that population substructure is the result of individuals migrating into a new area (i.e. dispersal). **BPEC** differs from the methods of Lemey *et al.* (2009) in that it explicitly models geographical ranges, assuming that sampling localities are random samples from the entire landscape. In contrast to the continuous approach of Guindon *et al.* (2016), it addresses the phylogeographic structure by inferring geographically structured clusters of DNA sequences as the result of distinct colonisation events, while also admitting a model for the evolutionary history. Here a cluster is defined as a subnetwork of sequences within the haplotype tree that are geographically aggregated and have similar ecological characteristics. **BPEC** performs full Bayesian inference, which means that it provides an entire posterior distribution over phylogeographic clusterings; although this comes at a computational cost, the ability to provide uncertainty measures is valuable in terms of understanding the impact on scientific hypotheses of interest.

The key function of **BPEC** inputs non-recombinant DNA sequences and geographical locations, as well as any additional covariates available, such as temperature or phenotypic characteristics, in order to identify clusters that are consistent with migration. The results of the analysis provide estimates on the number of migration events, the geographical distribution of the clusters, ancestral locations and clustered tree structure. Aside from providing estimates for the quantities of interest, **BPEC** also provides measures of uncertainty of the conclusions and functions for post-processing. Finally, **BPEC** is supplemented with various visualization tools interfacing with geographical mapping resources to aid interpretation. In Section 2, we present the **BPEC** model, followed by the corresponding Bayesian computation methods in Section 3. Section 4 describes an example dataset of the eastern lineages of polytypic Near Eastern brown frogs, *Rana macrocnemis* (Boulenger 1885), from the Caucasus region (Tarkhnishvili *et al.* 2001), and the R user interface is presented in Section 5 through the analysis of the example dataset. The output is interpreted in Section 6 and the paper concludes with a short discussion in Section 7.

2. Model

The aim of **BPEC** is to combine sequence data $\mathcal{S}$ with geographical and (optionally) ecological data $\mathcal{Y}$ for demographic inference regarding the geographic and (optionally) ecological structuring of genetic variation and thus potential geographical or ecological limitations to gene flow. To achieve this aim **BPEC** combines an evolutionary model for the genealogical relationships among sampled DNA sequences together with a geographical model representing dispersal events forming clusters into a fully model-based framework.

2.1. Haplotype tree model

Approaches to model and estimate the evolutionary relationships among DNA sequences range from simple and elegant, such as the vanilla coalescent (Kingman 1982) to complex with intractable likelihood forms (Cornuet *et al.* 2014). Questions such as the validity of a constant (or effectively constant) population size, independent nucleotide mutations, constant

mutation across sites, time-dependence, presence of natural selection pressure, all play a role in defining an appropriate evolutionary model and have led to a variety of extensions of the basic model (Wakeley 2013; Hein *et al.* 2004). In our case, typical datasets are expected to vary from a several hundred to no more than several thousand nucleotides, with low levels of polymorphism that typically characterise intraspecific data sets.

As a result, the nucleotide data are noisy and often too weakly informative to allow for very complex models. The evolutionary relationships among a sample of DNA sequences can be represented in one of two ways: a coalescent tree or a haplotype tree or network. A coalescent tree is plotted against time and thus explicitly characterizes the most recent common ancestor. An example of a coalescent tree with mutations mapped on is shown in Figure 1 where tips represent observed sequences and black circles indicate mutations, and the timing of the most recent common ancestors (MRCAs) among sequences is represented by branch lengths. In contrast, haplotype trees (Figure 2) summarise mutation differences among sampled sequences, so only implicitly carry information about time. To infer the root haplotype within a haplotype tree an evolutionary model is needed, but such models are not readily available. However, models such as the coalescent with mutations are available (Ethier and Griffiths 1987).

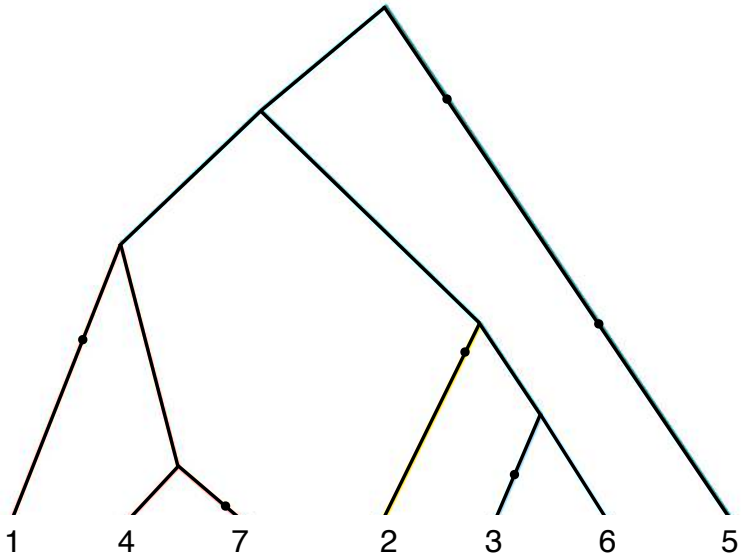

Figure 1: An example of a coalescent tree with mutations for 7 observed sequences including mutations; black dots represent mutations. Time evolves from top to bottom.

A subtle complication derives from potential tree unidentifiability due to repeated observations of the same haplotype that are to be expected when either or both the mutation rate and number of sampled nucleotides is insufficient to ascribe unique variation to all sample haplotypes. As an example, observations 4 and 6 in Figure 1 correspond to the same haplotype, meaning that the 2 observations could be switched without having any effect on the likelihood of the tree. Observations may, however, be distinct with respect to the geographical or ecological information associated to each one. Aside from identifiability issues, exploring the space of equivalent trees requires cycling through a complex combinatorial object which quickly becomes computationally cumbersome. Collapsing sequences into haplotypes allows

us to get around this issue, reducing the space of possible trees as exemplified in Figure 2 where sequences 4 and 6 from the coalescent tree (Figure 1) are now represented by the same node.

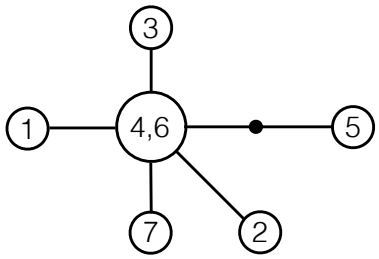

Figure 2: The corresponding haplotype tree of Figure 1, where edges represent single effective mutations. The black dot represents an unobserved intermediate sequence. Note that sequences 4 and 6 correspond to the same haplotype.

In order to draw inferences about the haplotype tree, approaches can be fully model-based (Felsenstein 1983; Huelsenbeck and Ronquist 2001; Drummond *et al.* 2012), parsimony-based through the underlying tree (Rzhetsky and Nei 1993; Desper and Gascuel 2002), or purely phenetic such as neighbour-joining or median-joining (Atteson 1999; Gascuel and Steel 2006). **BPEC** combines parsimonious approaches within a model-based framework. Although an infinite set of haplotype or coalescent trees could be consistent with the sequence data $\mathcal{S}$, **BPEC** uses relaxed parsimony to reduce it to a finite set of 'plausible' trees Ω represented via a graph (Manolopoulou and Emerson 2012). The relaxed parsimony is defined by a threshold d_s representing parsimony relaxation. Briefly, haplotypes are connected by an edge if they are a single mutation apart. When two groups of haplotypes are disconnected (with minimum mutation distance d_{min}), then any connection path with length up to $d_{min} + d_s$ is considered. The exact details of how to obtain Ω from $\mathcal{S}$ for a given d_s can be found in algorithm A of Manolopoulou and Emerson (2012). This algorithm constructs a set of 'realistic' trees by cumulatively adding intermediate sequences following a relaxed parsimony assumption defined by the user-specified parsimony relaxation parameter d_s. In general, larger values of d_s (up to a maximum value) yield more inclusive (and hence realistic) sets Ω, but the choice of d_s is often limited by computational power. For a fixed d_s, this algorithm inputs the DNA sequences at hand, and outputs a sequence network, including loops. The true haplotype tree is then assumed to be one of the minimum spanning trees of this graph with equal probability and can be obtained through the breaking of loops.

A haplotype tree encodes less information than a coalescent tree with mutations. Firstly, a haplotype tree only encodes time through number of mutations. Secondly, it does not automatically define an ordering of events, starting from a root down to tips. Even a rooted (i.e., one where the ancestral haplotype is specified) haplotype tree imposes only a partial ordering to the set of past mutation and coalescence events. Calculating probabilities over rooted haplotype trees therefore requires summing over all possibilities and orderings of past events given a temporal model; an example of possible orderings is shown in Appendix 8.1. We denote a temporal ordering of events as $\mathcal{O}$, where $\mathcal{O}^{\mathcal{S}}_{r,T}$ denotes the set of all temporal orderings consistent with data $\mathcal{S}$ given a root r and tree T and assume that any temporal ordering of events is equally likely a priori. Conditionally on observed data (which restricts the possible trees to the space Ω) this prior corresponds to a discrete uniform distribution

over Ω and provides the following posterior probabilities for the root r and tree T:

$$\mathbb{P}(r, T \mid \mathcal{S}) = \frac{\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}{\sum_{r,T}\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|},$$

where $|\cdot|$ denotes the size of the set. Similarly,

$$\mathbb{P}(r \mid T, \mathcal{S}) = \frac{\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}{\sum_{r}\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|},$$

$$\mathbb{P}(T \mid r, \mathcal{S}) = \frac{\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}{\sum_{T}\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}.$$

This model naturally takes into account the total number of combinations of mutational and coalescence events. Note that this model disregards the relative probability of coalescence versus mutation, essentially assuming that at every time point either are equally likely. The model can be extended to introduce a mutation rate θ (at the expense of computational complexity) which is simultaneously learnt and is used to refine the posterior probabilities of each tree. Although the haplotype tree model described provides a way of assigning posterior probabilities of haplotypes being ancestral, these need to be associated to sampling locations in order to infer the most ancestral location. **BPEC** assigns probabilities to each location based on the haplotypes observed in each. For each posterior sample, if the inferred root haplotype is observed, then each observed sequence that corresponds to that haplotype contributes equally to a location being ancestral. In other words, each location will be inferred to be ancestral with probability equal to the proportion of root haplotypes that were sampled in it. If the inferred root haplotype is not observed (i.e. extinct or unsampled), then the oldest observed haplotypes derived from the inferred root are considered equally likely to be the 'most ancestral' and thus each observation of one of these haplotypes contributes equally to the probability of each sampling location being ancestral. An example of this is shown in Figure 3. An important feature of this approach is that the probability of each location being ancestral depends on the proportional representation of each haplotype. This is to circumvent issues of wide sampling variability across locations.

2.2. Clustering model

The two main requirements to infer migration events for a given tree are: (i) a model for constructing constrained clusterings conditionally on a haplotype tree, and (ii) a model for the distribution of data within each cluster. A key assumption in our model is that new clusters are formed through the migration/dispersal/colonisation of a single individual (haplotype) founding a new geographically distinct cluster (De Iorio and Griffiths 2004a,b). All subsequent descendants of this founding haplotype belong to the new cluster, unless they migrate again. Given an inferred tree representing the genealogy, possible clusterings of the data are thus constrained by the tree while at the same time informed via the geographic distribution (and optionally ecological data) of the observations for each individual. Figure 4 provides an illustration using the hypothetical coalescent tree of Figure 1.

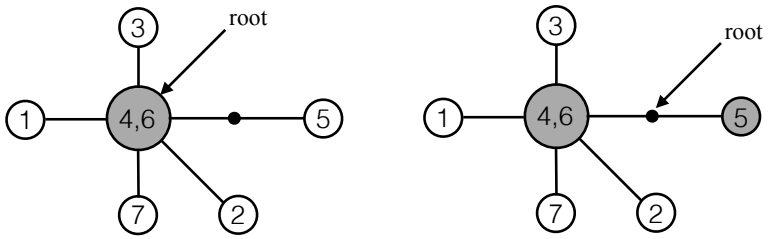

Figure 3: Two possible root scenarios for the haplotype tree presented in Figure 2. In the left-hand panel, the root haplotype (shown in grey) is observed and thus any location will be inferred to be ancestral according to the proportion of observations of the inferred root haplotype 4. In the right-hand panel, the inferred root haplotype is not observed and the two equally divergent descendant haplotypes (shown in grey) are then used to infer ancestral locations as a function of the proportion of copies of either haplotype in a given location.

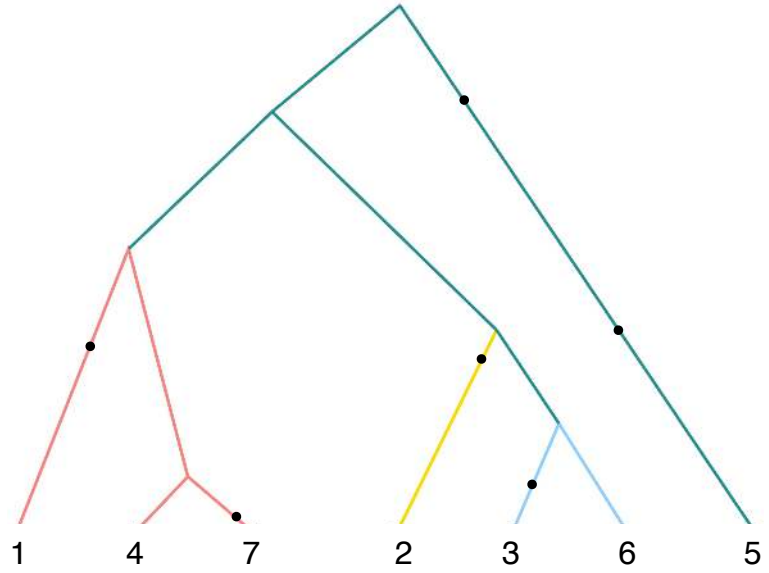

Figure 4: Geographically informed genetic clustering of the coalescent tree from Figure 1. Different geographic clusters are represented by different colours. The inferred ancestral geographic area is represented by green with three migration events inferred to give rise to three derived geographic clusters.

The coalescent tree determines a set of constrained clusterings which are feasible through migration events. For example, observation 2, 3 and 6 in Figure 4 could have formed a single cluster together, but 6 and 7 could not. The corresponding constrained clusterings defined on the collapsed haplotype tree are slightly less intuitive as repeated observations of the same haplotype (node) can belong to different clusters. In the collapsed haplotype network shown in Figure 5, all clustered nodes must be directly connected within their cluster. For simplicity, we shall refer to haplotype 4/6 as haplotype 4 from now on.

Formally, conditionally on a haplotype tree, the clustering model is defined as follows. We denote the set of distinct haplotypes in the sequence set $\mathcal{S}$ (of size N) as $\mathcal{H} = \{H_1, \ldots, H_n\}$ with size n, and use $|H_i|$ to denote the number of copies of haplotype H_i observed in the data.

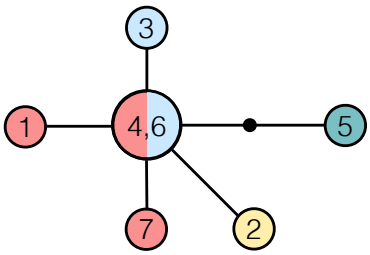

Figure 5: The clustered haplotype tree corresponding to the subdivided coalescent tree of Figure 4 where colour corresponds to cluster and size of node to the number of individuals sampled with each sequence. Edges represent single effective mutations and black dots represent unobserved intermediate haplotypes.

Let K denote the number of migrations, which is itself allowed to vary. Each migration event is associated with a haplotype which migrated, denoted as $\mathbf{m} = \{m_1, \ldots, m_K\}$. Although colonisation events happen in order, here we do not model the events temporally, so the order of $\mathbf{m}$ is irrelevant. Note that the haplotypes in this list need not be distinct, as two different copies of the same haplotype may have colonised, or a single sequence may have colonised twice. The set of colonies/clusters with which each migrating haplotype is associated is denoted by $\mathcal{C}(m_k),\ k = 1, \ldots, K$; in the example above, $K = 3$, $\mathbf{m} = \{4, 4, 4\}$ and $\mathcal{C}(4) = \{\text{blue,yellow,pink,green}\}$, since all migrations were of the same haplotype. This means that, in general, the sample space of $\mathbf{m}$ has size $n^K/K!$.

Conditionally on a set of migrating haplotypes, the space of constrained clusterings is then such that all observations of that haplotype must belong to one of the corresponding clusters $\mathcal{C}(m_k)$ (i.e., either the original cluster or the one which was a result of migration). Equivalently, all adjacent haplotypes must also belong to one of these clusters (unless one of them also migrated, and so on).

Once the clustering has been established, the geographical and ecological observations $Y_i,\ \ i = 1, \ldots, N$ in each cluster c_i are Normally distributed with mean μ_c and variance Σ_c, such that

$$Y_i \ \sim \ N(\mu_{c_i}, \Sigma_{c_i}),\ \ i = 1, \ldots, N, \tag{1}$$

where c_i denotes the cluster of observation i.

To complete the model, prior distributions are defined on the model parameters. The number of migrations is assumed to be uniform between 0 and $K_{\max}$ (corresponding to 1 and $K_{\max}+1$ clusters). Other prior distributions (e.g., Poisson) could be used instead, but we do not explore this direction here. The $|m_k|$ observations of each of the migrating haplotypes m_k are each assigned uniformly to one of the clusters in $\mathcal{C}_k$, similarly with the $deg(m_k)$ clades connected to it (where degree represents the number of edges connected to node m_k), so the prior probability of each clustering conditional of the migrating haplotypes (and their clusters) is simply a combinatorial coefficient.

$$\begin{aligned} K \ &\sim \mathcal{U}\{0, \ldots, K_{\max}\}, \\ \mathbf{m} \mid \mathcal{S} \ &\sim \text{Multinomial}\{|H_1|, \ldots, |H_n|\} \\ \text{and } p(\mathbf{c} \mid \mathbf{m}, T) &= \prod_{k=1}^{K} \left(\frac{1}{|\mathcal{C}_k|}\right)^{|m_k|+deg(m_k)} \end{aligned} \tag{2}$$

The means and variances of each clustering are assigned different priors for the longitude-latitude versus the remaining covariates:

$$\begin{aligned} \Sigma_{k,(1:2,1:2)} &\sim \mathcal{IW}(\gamma, \psi\mathbb{I}_2),\ k = 1, \ldots, (K_{\max}+1), \\ \Sigma_{k,(3:d)} &\sim \mathcal{IG}(\gamma, \psi),\ k = 1, \ldots, (K_{\max}+1), \\ \gamma &\sim \mathcal{U}\{4, \ldots, g\}, \\ \mu_k \,|\, \Sigma_k &\sim \mathcal{N}(\mathbf{0}, V),\ k = 1, \ldots, (K_{\max}+1), \end{aligned} \tag{3}$$

and we assume that any off-diagonal entries of Σ_k in dimensions $3:d$ are 0. By convention, the first two coordinates of Y always represent longitude and latitude, normalised such that the mean of both is zero and the average (between longitude and latitude) variance 1, using the same normalising factor for both longitude and latitude to reflect the isotropy of the two dimensions. Note that longitude and latitude are treated as Euclidean coordinates, which means that datasets spanning a very large region may result in distorted results. The remaining coordinates correspond to environmental or phenotypic characteristics (if available), which are normalised to sample mean 0 and marginal variance 1. We impose uncorrelated environmental/phenotypic characteristics by forcing the covariance matrices to be 0 on any off-diagonal entries except for the one correponding to longitude-latitude. This is because the concentration parameter γ of an Inverse Wishart needs to be at least as $d_\Sigma + 2$ in order to be well-defined, where d_Σ is the dimension of the covariance matrix modelled. In our case, if we model the entire covariance matrix through an Inverse Wishart, γ would be forced to a minimum of $3 + d$, which (for moderate d) corresponds to low prior variance and can be too restrictive. We thus restrict the Inverse-Wishart prior for the geographical covariates only and place independent Inverse-Gamma priors on the remaining diagonal elements of Σ_k.

Perhaps the most important prior distributions here are the ones relating to the shape Σ of each cluster, namely the parameters of the Inverse-Wishart prior γ and ψ, as these define the prior belief of the spread of each cluster. Although the parameter γ is allowed to vary and hence can adapt depending on information from the data, nevertheless too large or too small values of ψ (corresponding to a prior belief of geographically widely spread versus tiny clusters) will have an impact on the posterior inference. The default setting in **BPEC** is that clusters are a priori expected to span about 30% of the total range.

3. Bayesian computation

The entire model consists of the model of the root and tree posterior distribution together with the distribution of the migration and clustering model. Inferences are drawn simultaneously, such that we can borrow information from the tree to the migration parameters and vice versa. The complexity of this phylogeographic model implies that drawing inferences about the posterior distribution of the parameters is challenging. We proceed via tailored Markov chain Monte Carlo (MCMC) using a combination of adaptive proposals, auxiliary variables and data-driven proposals. This is especially crucial for the clustering, which here is restricted to tree-based clusterings, since the space of clusterings is vast and discrete without natural local moves.

3.1. Markov chain Monte Carlo sampler

The Markov chain Monte Carlo sampler alternates between updates of the tree parameters and

the clustering parameters. We adopt a scheme whereby updates of parameters are performed at varying frequencies, reflecting the difficulty of accepting or rejecting a move and allowing both local and global exploration of the parameter space. Four different updates are described below, which are then combined into a sampler at varying frequencies.

The tree T, root r, colonised haplotypes $\mathbf{m}$, clustering $\mathbf{c}$ and cluster means $\boldsymbol{\mu}$ and variances $\boldsymbol{\Sigma}$.

1. Conditionally on a given tree T, propose to change the root along with a mutation history. Accept or reject the proposed root and mutation history.

2a. Conditionally on the root r, propose a new tree T and mutation history uniformly.

2b. Conditionally on the proposed tree T propose to change one of the colonised haplotypes in $\mathbf{m}$.

2c. Conditionally on the colonised haplotypes, propose to change the set of clusterings $\mathbf{c}$ along with the means $\boldsymbol{\mu}$ and variances $\boldsymbol{\Sigma}$ of each cluster.

2d. The proposed tree topology and history, root, clustering and means and variances are accepted or rejected together. However, steps (2a), (2b) and (2c) need not all occur at the same time. Specifically, steps (2a-b) are only performed (roughly) every 5th iteration.

3. Conditionally on a given clustering, update the cluster means conditionally on all other parameters, and subsequently the sample covariance conditionally on all other parameters.

4. Propose to increase or decrease the number of clusters. Then propose to add or subtract a colonised sequence, then set of clusterings together with means and variances of each cluster. Accept or reject the entire move.

The precise mechanics of the sampler are not shown here; some additional technical issues are discussed in Appendix 8.2.

3.2. Technical considerations

Almost as important as how the method works is when it is sound to use (or not). Since the package is intended to be used primarily by practitioners, one of the aims of this paper is to clarify what types questions **BPEC** can potentially answer as well as what underlying assumptions are necessary and implicit.

Bayesian Phylogeographic and Ecological Clustering assumes that non-recombinant (typically mtDNA) data are available from a set of geographical locations (in the form of longitude/latitude). The haplotype tree model takes a relaxed parsimony approach which may be unreliable under conditions of mutational saturation or excessive homoplasy. **BPEC** is programmed to produce appropriate error messages to inform the user in such cases, but will not be foolproof.

The geographical model assumes a constant population size and migration rate, and thus as real data departs from this model the inferences from **BPEC** are expected to depart from the true demographic history. However, simulation analyses will be required to address this

quantitatively. Also, the clustering and migration model does not explicitly take into account geographical distance between clusters. It simply separates observations in distinct geographical clusters. Therefore, it is possible for a migration to result in two distant clusters.

Notice that we assume a uniform prior over the number of migrations K. In general, K migrations can lead to up to $K+1$ clusters; often, however, some of these may be empty, resulting in fewer 'effective' migrations. The uniform prior applies to the total number of migrations rather than the number of effective ones, whereas the posterior distribution over the number of migrations actually refers to effective migrations. This somewhat convoluted approach is preferred because enumerating scenarios of different effective migrations is computationally cumbersome.

As discussed earlier, an important consideration when using **BPEC** for the inference of ancestral areas is the distribution of haplotype observations within each location. Since ancestral area probabilities are determined through the proportion of inferred ancestral haplotypes, a site with, for example, a single haplotype which happens to be ancestral, will always result in high probability of being ancestral. Consequently, ancestral location probabilities should be more reliable when there are more observations per location. It also frequently occurs that uncertainty about the root haplotype is high, where a range of different haplotypes carry significant posterior mass. As long as no convergence errors are reported, this is not a convergence issue but merely reflects uncertainty in the data.

One of the limitations of Markov chain Monte Carlo methods is that the samplers require a large number of iterations to satisfy convergence diagnostics. The convergence diagnostics in **BPEC** are split into two pieces: convergence of the clustering and convergence of the root haplotype. If either of these two pieces has not converged, the sampler will return an error to that effect. Ideally, both pieces should satisfy the convergence diagnostics; however, it is sometimes the case (especially when dealing with a large number of clusters) that, for any reasonable number of MCMC iterations, the diagnostics fail. In these cases, inferences should be taken with caution.

BPEC cannot deal with unknown nucleotides and will ignore any nucleotide sites at which at least one of the sequences has an ambiguity code. This means that ambiguous nucleotides result in information loss. On the other hand, **BPEC** will treat true alignment gaps '-' as a 5th character such that a deletion/insertion is treated as a type of mutation. Care should be taken in the interpretation of the output when lots of missing nucleotides are present, since this could lead to significant loss of resolution (Joly *et al.* 2007).

4. Brown frog data

The **BPEC** package will be implemented on a brown frog dataset which will be used throughout the next few sections for illustration. We used 40 mitochondrial cytochrome b sequences of Near Eastern brown frogs - *Rana macrocnemis* (Boulenger, 1885) to demonstrate a combined phylogeographic and ecological analysis with **BPEC**. Previous molecular analyses (Tarkhnishvili *et al.* 2001; Veith *et al.* 2003b,a) have attributed range expansion and fragmentation triggered by Pleistocene glaciation cycles as drivers of demographic change within the brown frog.

R. macrocnemis is represented by a number of recognised subspecies across its entire range, and here we focus on two widespread subspecies that are geographically distinct in the south-

west Caucasus and separated by a narrow transition zone (Tarkhnishvili *et al.* 2001). The nominotypic *R. macrocnemis macrocnemis* (Boulenger, 1885) is found on the forested slopes of the Trialeti ridge northwest and in montane meadows on both sides of the Great Caucasus, while *R. macrocnemis camerani* (Boulenger, 1885) occurs in southern Georgia on the Javakheti plateau (Tarkhnishvili *et al.* 2001). A map of the sampling localities, indicating proportion of *R. macrocnemis macrocnemis* versus *R. macrocnemis camerani*, is shown in Figure 6.

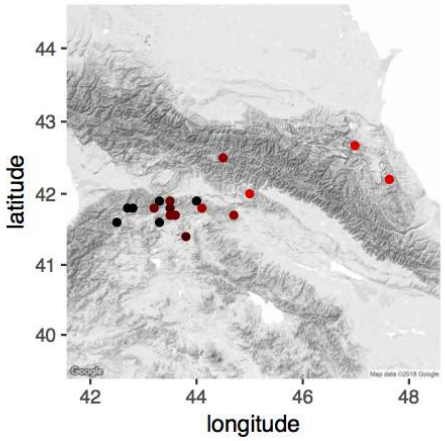

Figure 6: The sampling localities for *R. m. camerani* and *R. m. macrocnemis* overlayed onto the map. Colour represents proportion of *R. m. camerani* versus *R. m. macrocnemis* individuals sampled, with red corresponding to 100% *R. m. camerani*, black 100% *R. m. macrocnemis* and brown corresponding to mixed populations.

BPEC was applied to investigate geographic and environmental aggregation of haplotypes within *R. macrocnemis*. We included predictive environmental and climate covariates (topographic and land cover conditions, and annual trend patterns of temperature, precipitation and seasonality) to examine environment and geography as agents for the structuring of genetic variation. Grid-based attribute values of a set of predictor variables associated with each cell position of the map layers were subsequently extracted at the point locations of the georeferenced mtDNA haplotypes from six raster grids by means of the `extract` function of the **raster** package (Hijmans 2019): four bioclimatic variables (Annual Mean Temperature (degrees Celsius x 10), Temperature Annual Range (100 × standard deviation of monthly mean temperature), Annual Precipitation (in mm), Precipitation Seasonality (Coefficient of Variation (CV)), from the bioclim database available in the **dismo** package (Hijmans *et al.* 2017), altitude in meters as a proxy for a digital elevation model and the land cover map (GLC2000, Bartholome and Belward (2005)) from the subdomain land cover/land use housed under `http://worldgrids.org/` global environmental layers. We re-classified the total information of the land cover map into two classes of forested and non-forested areas to introduce a simplistic landscape dependent habitat variable (COV). These six variables altogether describe climatic, topographic, and land cover conditions that are potentially informative predictors in terms of species distribution.

5. User interface

5.1. Inputs

BPEC takes two main inputs: the set of mtDNA sequences (in NEXUS format) and the set of coordinates and haplotypes observed in each location. Sequences need not be collapsed into unique haplotypes, but labelling of sequences in the NEXUS file and the locations file must be consistent. In order to load these two variables into R from two files called `haplotypes.nex` and `coordsLocsFile.txt` (for example), the following commands can be used. For an example of input files, use the files provided through `system.file("haplotypes.nex",package = "BPEC")` and `system.file("coordsLocsFile.txt",package = "BPEC")` or see Supplementary materials of the manuscript.

The sequences can be loaded using the `bpec.loadSeq` command.

```
R> library("BPEC")
R> rawSeqs <- bpec.loadSeq("haplotypes.nex")
```

The file `coordsLocsFile.txt`, containing the list of coordinates, covariates and haplotypes, needs to have (in each row): latitude, longitude, environmental/phenotypic covariate values (if available), plus a set of numbers corresponding to the haplotypes/sequences with these attributes. For example, for two locations with 5 observations in total from 3 haplotypes, with no additional covariates, the file might read

```
40.3 45.2    1    2    2
45.3 50.1    2    3
```

All haplotypes/sequences found in a location can be entered in one line, or only one per row, such that we could also have used

```
40.3 45.2    1    2
40.3 45.2    2
45.3 50.1    2    3
```

or any such combination.

When additional environmental or phenotypic covariates are available, these can also be entered as a column right after the longitude and latitude, such as

```
40.3 45.2 18.1    1
40.3 45.2 22.5    2    2
45.3 50.1 25.0    2    3
```

where $18.1, 22.5, 25.0$ are, for example, temperatures. Names for the covariates at each location can optionally be provided through the header row using the option `header = TRUE`, these will later appear in the output plots to aid interpretation.

```
lon  lat  temp
40.3 45.2 18.1    1
40.3 45.2 22.5    2    2
45.3 50.1 25.0    2    3
```

Environmental covariates can be extracted, for example, from publicly available databases such as bioclim by means of the R package **raster** (Hijmans 2019).

In order to load the file containing the coordinates, covariates and observed haplotypes/sequences of each location, use the `bpec.loadCoords` command below. Use the option `header = TRUE` when the first row includes variable names.

```
R> coordsLocs <- bpec.loadCoords("coordsLocsFile.txt", header = TRUE)
```

The brown frog dataset is in-built and can be loaded through

```
R> data("MacrocnemisRawSeqs")
R> data("MacrocnemisCoordsLocs")
R> rawSeqs <- MacrocnemisRawSeqs
R> coordsLocs <- MacrocnemisCoordsLocs
```

which contain the 40 sequences together with their corresponding longitude/latitude, along with 6 environmental covariates. Other datasets that are available in **BPEC** can be found using `data(package = "BPEC")`.

5.2. Main MCMC command and options

Once the `rawSeqs` and `coordsLocs` variables have been loaded, the Markov chain Monte Carlo sampler can be run through the command

```
R> bpecout <- bpec.mcmc(rawSeqs, coordsLocs, maxMig = 3, iter = 1000000,
+                       ds = 3, postSamples = 1000, dims = 8)
```

The arguments are described in Table 1.

In the case of the brown frog dataset, the dimensionality of the data was `dims = 8` (geographical dimensions longitude and latitude plus the six additional environmental covariates). We ran the **BPEC** analysis taking the maximum parsimony level option at `ds = 0`, increasing up to `ds = 3` (to potentially explore more candidate trees) for 1,000,000 iterations (`iter`) each. No change to the results was observed, since the brown frog haplotypes formed a fully connected tree without missing intermediate haplotypes. Convergence diagnostics of the maximum a posteriori clusterings and root were not violated (i.e., no convergence error message was reported). The output of the function is shown below.

```
Starting bpec...
Inferring possible missing sequences....
Counting loops in the network...

The program found no loops that need to be resolved in the network

Number of iterations is 1000000
Number of saved iterations 1000
Sample size is 40
Effective sequence length is 8
```

`bpec.mcmc()`	
Argument	*Description*
`maxMig`	the maximum number of migrations to be considered. In terms of inference, the higher `maxMig`, the better the results, since more models are considered. However, that comes at a computational cost. We recommend using a low but intuitive value based on the study system to begin an iterative assessment. For example, if using a value of 6 (corresponding to 7 clusters), and if the inference shows significant posterior probability on 7 clusters, increase `maxMig` and re-run. Similarly, if e.g., a value of 5 is used and convergence diagnostics are not satisfied, but posterior mass seems to be minimal around 4/5 migrations, then one can reduce `maxMig` to 4 (which will reduce complexity) and re-run.
`iter`	the number of MCMC iterations to run the sampler for. By default, two chains will be run from different starting values. The value of `iter` is important, as it will determine how long the chains will run for and whether convergence (both in terms of the root haplotype as well as the clustering) diagnostics will be satisfied. A value of 100,000 is usually reasonable to start with; if convergence diagnostics are not satisfied, or if the post-processing plots look inconsistent, increase `iter` by a factor of 10 (and so on).
`ds`	the parsimony relaxation parameter d_s. We recommend starting with $d_s = 0$ and increasing once reasonable values of `iter` and `maxMig` have been established. Note that increasing d_s past an (unknown) value d_{max}, which depends on the individual dataset, has no effect on the inference.
`postSamples`	the number of posterior samples (per chain) to be saved for posterior summary statistics. We recommend using a value around 1,000. The higher the better for inference, but this comes at a memory storage cost.
`dims`	the number of covariates (including longitude and latitude) available. If only geographical data are used (and no environmental or phenotypic information), `dims=2`. Otherwise increase as appropriate.

Table 1: The list of inputs required to use `bpec.mcmc()`.

input()	
Output	*Description*
seqCountOrig	the number of sequences in the data.
seqLengthOrig	the length of the input sequences.
iter	the number of MCMC iterations.
ds	the parsimony relaxation parameter.
coordsLocs	the input coordinates (and optional additional ecological measurements) and their corresponding sequence indices.
coordsDims	the dimension of the input measurements (2 if purely longitude and latitude, +1 for every additional one).
locNo	the number of distinct sampling locations.
locData	the coordinates and measurements of each sampled sequence.

Table 2: The list of outputs of `input()`, corresponding to all the inputs and arguments that were provided to `bpec.mcmc()`.

```
Total number of haplotypes (including missing) 10
Dimension is 8
Parsimony relaxation is 3
Maximum number of migrations is 3

Starting MCMC sampler (burn-in ends at 90% and acceptance rate re-started):
Chain 1: |====================|100% (accepted samples 2763 time 24 minutes)
Chain 2: |====================|100% (accepted samples 2823 time 49 minutes)

The most likely root node is 2
The most likely ancestral locations are 38,34,4
```

5.3. Outputs

The `bpec.mcmc` command outputs an R object of class `BPEC` which can be summarised using generic functions such as `plot()`, `summary()` and `plot()`, as well as accessor functions `input()`, `preproc()`, `output.tree()`, `output.clust()`, `output.mcmc()`. The output of each of these accessor functions is shown in Tables 2-6.

5.4. Visualizations and post-processing

As described in the previous section, the Markov chain Monte Carlo sampler returns many different types of outputs. In order to obtain a summarised picture of the inference, a number of visualizations are available through **BPEC** to aid interpretation.

Geographical contour plot

The command `bpec.contourPlot` provides a colour-coded contour plot of the geographical clusters superimposed onto a map (provided accurate longitude and latitude coordinates have been provided) using

preproc()	
Output	*Description*
`seq`	The output DNA sequences of distinct haplotypes, collapsed to effective nucleotide sites (both sampled and missing sequences which were inferred).
`seqsFile`	A vector of the numerical labels of each haplotype.
`seqLabels`	Correspondence vector for each of the processed observations to the original haplotype labels.
`seqIndices`	Correspondence vector for each of the original observations to the resulting haplotype labels.
`seqLength`	The effective length of the input sequences, given by the number of variable nucleotide sites which are informative. In other words, if two or more nucleotide sites describe the same subsets of sequences, then they are collapsed to a single informative nucleotide.
`noSamples`	The number of times each haplotype was observed in the sample.
`count`	The number of output sequences.

Table 3: The list of outputs of `preproc()`, corresponding to values arising from the data before the Bayesian analysis.

output.tree()	
Output	*Description*
`clado`	the adjacency matrix for the maximum a posteriori tree in vectorised format. For two haplotypes `i,j`, the `(i,j)`th entry of the matrix is 1 if the haplotypes are connected in the network and 0 otherwise.
`levels`	Starting from the root (level 0) all the way to the tips, the discrete depth for the maximum a posteriori tree.
`edgeTotalProb`	Posterior probabilities of each edge being present in the tree, so that any edge which is not part of a loop will have posterior probability 1.
`rootProbs`	a vector of the posterior probabilities that each haplotype is the root of the tree.
`treeEdges`	contains the same information as `cladoR`, but in a different format. The set of edges (from and to haplotypes) of the maximum a posteriori haplotype tree are represented as an edge list of from/to vectors which could be used in the graph and network modelling R package **igraph** (Csardi and Nepusz 2006) if needed.
`rootLocProbs`	a vector of the posterior probabilities of each sampling location being the most ancestral location. If several rows in the file `coordsLocsFile.txt` correspond to the same geographical location, the first of these will carry the total posterior probability for the location, with the remaining having 0.
`migProbs`	a vector of the posterior probabilities of {`0...maxMig`} migrations.

Table 4: The list of outputs of `output.tree()`, corresponding to the output of the tree model.

output.clust()	
Output	*Description*
sampleMeans	a set of postSamples posterior samples of the cluster centres.
sampleCovs	a set of postSamples posterior samples of the cluster covariances.
sampleIndices	a set of posterior samples of the cluster allocations of each observation.
clusterProbs	for each haplotype, posterior probabilities that it belongs to each cluster.

Table 5: The list of outputs of output.clust(), corresponding to the output of the geographical clustering model.

output.mcmc()	
Output	*Description*
MCMCparams	various tuning parameters used in the MCMC sampler, this is only important for development.
codaInput	Posterior samples from the two MCMC chains for the cluster means, cluster covariance entries, as well as the root haplotype. Note that, since the number of clusters varies from iteration to iteration, some samples are simply draws from the prior (corresponding to empty clusters). This variable can be loaded directly into the **coda** package (Plummer *et al.* 2006) for convergence analysis.

Table 6: The list of outputs of output.mcmc(), corresponding to technical MCMC aspects.

```
R> par(mar = c(0, 0, 0, 0))
R> bpec.contourPlot(bpecout, GoogleEarth = 0, mapType = "osm",
+                   colorCode = c(7, 5, 6, 3, 2), mapCentre = NULL, zoom = 7)
```

In order to convey not only posterior means but also uncertainty, a set of posterior draws of these contours are plotted using transparency, so that the user can assess the stability of the inference.

The sampling locations are also shown on this contour plot, with the top three sampling locations in terms of their probability of being ancestral shown as larger points. The precise posterior probabilities (which may all be low in the presence of uncertainty) of each of the localities being ancestral can be found through output.tree(bpecout)$rootLocProbs.

The colours can be changed through the optional argument colorCode (with default value (7,5,6,3,2,8,4,9)) which controls the colour of the first, second, third cluster etc; if not specified, the default colour scheme is used. There are four options for the argument mapType: 'none' will show the posterior distribution of the clusters against a white background, 'plain' will use the in-built outline R maps, 'google' will superimpose the contours on a map downloaded from Google maps (requires an internet connection) Kahle and Wickham (2013), and 'osm' will do the same using OpenStreetMap. The optional arguments mapCentre and zoom allow the user to specify the centre of the map and level of zooming when using the Google maps option.

In the case of the brown frog dataset, the contour plot is shown in Figure 7. The posterior mass for the number of clusters strongly concentrates around 2 (as indicated by the output

`output.clust(bpecout)$migProbs`), with the posterior probability of 2 clusters being greater than 0.99. The yellow cluster can be taxonomically aligned to the subspecies *R. m. macrocnemis* lineage, while the turquoise cluster includes individuals of *R. m. macrocnemis* from the humid and forested mountain region, and individuals assigned to *R. m. camerani* from the drier area of the southern treeless mountain steppe habitats of the Javakheti plateau. The contour ellipses overlap in the heart of the geographic transition zone south of the Minor Caucasus.

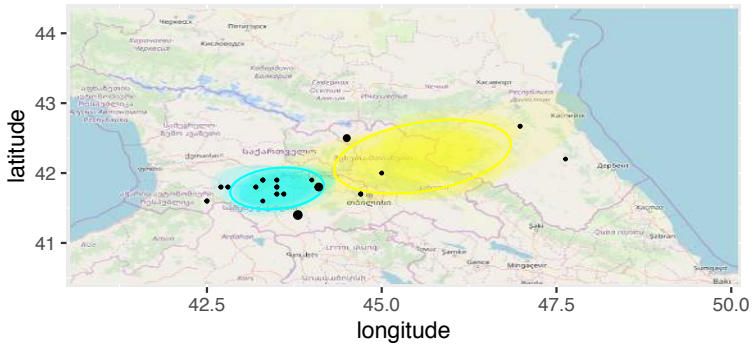

Figure 7: An example of the contour plot for the brown frog dataset using `bpec.contourPlot`. Each transparent geographical ellipse represents a posterior draw for the geographic centre of the cluster within the 50% level contour of that draw. The 50% contour represents the boundary where probability density of the cluster is 50% of the maximum density (i.e., the centre of the cluster). Solid ellipses represent posterior means. Larger triangles represent most likely ancestral locations. The black jagged lines show the outline of the geographical map of the area.

Instead of using the R interface, the contour plot can also be exported into Google Earth primary exchange format using the option `GoogleEarth = 1`. This will produce a set of files with extension `kml` which can be loaded directly into Google Earth.

Finally, a 'messy' looking plot such as the toy example in Figure 8 either implies poor MCMC convergence or high uncertainty in terms of the clustering.

Environmental and/or phenotypic covariates plot

In cases where environmental or phenotypic covariates have also been used, posterior draws for the distribution of the covariates within clusters are available through cluster means `output.clust(bpecout)$sampleMeans` and covariances `output.clust(bpecout)$sampleCovs`. These can be summarized through posterior medians and 5/95% credible regions, colour-coded using the same coding as the contour plot. To aid plotting and interpretation, the covariate names of each of the columns of `coordsLocs` are used. The first two (corresponding to longitude and latitude) are automatically ignored in this function.

```
R> par(mfrow = c(2, 3))
R> bpec.covariatesPlot(bpecout, colorCode = c(7, 5, 6, 3, 2))
```

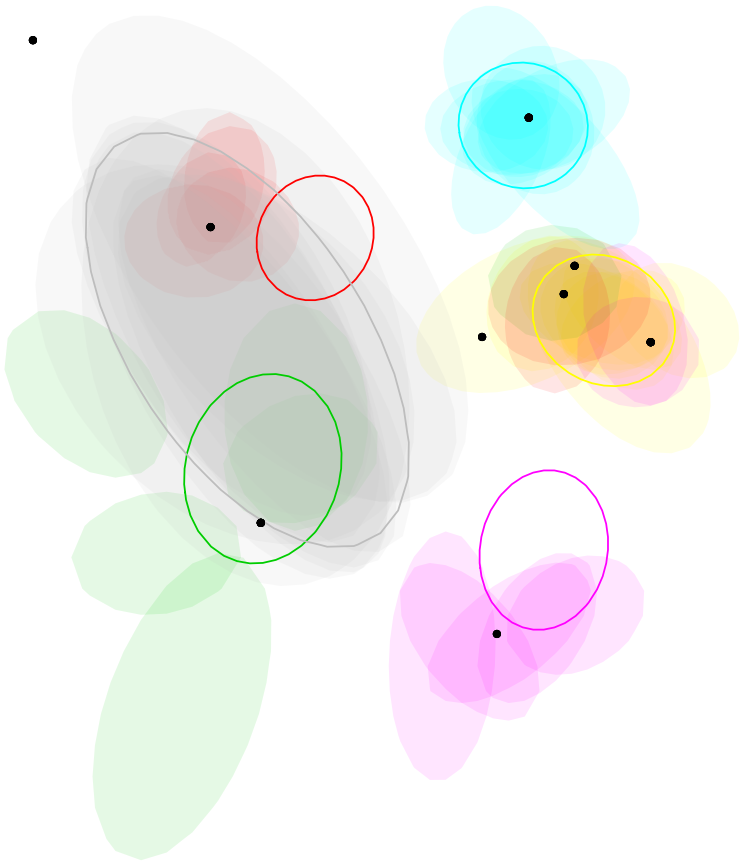

Figure 8: A toy example of a posterior distribution of phylogeographic clustering which shows high uncertainty. Many of the clusters 'jump' from one location to another from iteration to iteration, indicating uncertainty about the location (and number) of clusters.

The plot produced in the case of the brown frog dataset is shown in Figure 9.

Clustered tree plot

To visualize the maximum a posteriori haplotype tree, the command `bpec.treePlot` plots the haplotype tree most supported by the data. The size of each node in the tree represents the number of times each haplotype was observed, black dots corresponding to missing intermediate haplotypes. The thickness of each edge represents the posterior probability that each mutation occurred (thin edges corresponding to mutations with high uncertainty). Observed haplotypes are colour-coded according to their posterior probability of belonging to each cluster. As long as the same `colorCode` variable is used, the cluster colours correspond to the ones used in the geographical and covariate contour plots.

```
R> bpec.tree <- bpec.treePlot(bpecout, colorCode = c(7, 5, 6, 3, 2))
```

The corresponding plot for the brown frog dataset is shown in Figure 10. **BPEC** collapsed sequences to 10 distinct haplotypes with effective length 8 displayed in the maximum a posteriori haplotype tree shown.

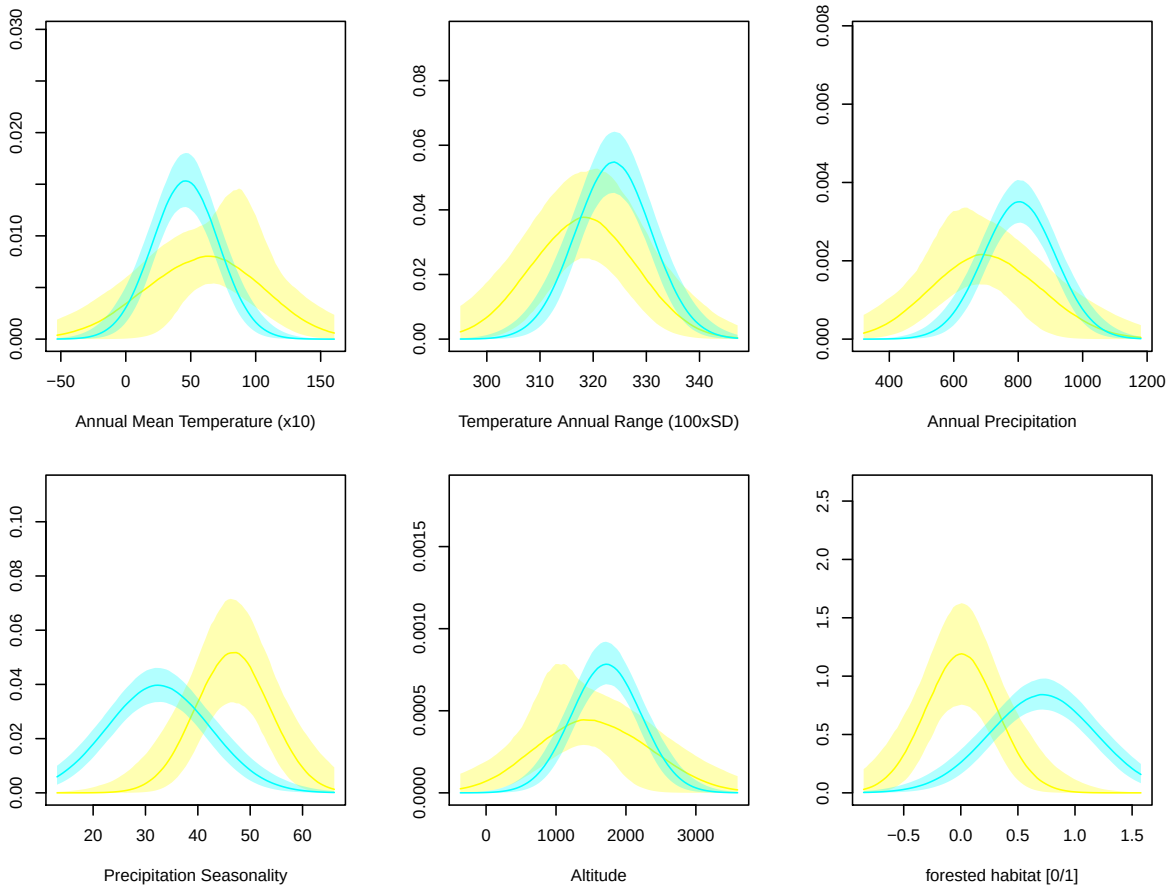

Figure 9: Plot of the distribution of the covariates for each cluster for the brown frog dataset using `bpec.covariatesPlot`. Shaded regions correspond to 5% and 95% pointwise credibility bands of each cluster, with solid lines showing the pointwise median. Colour corresponds to the same clusters as the contour plot above.

Tree plot on geographical map

The tree plot can also be partially visualised geographically through the `bpec.geoTree` command which superimposes the haplotype tree onto a map through a file that can be loaded into Google Earth. The function uses the **igraph** package (Csardi and Nepusz 2006) as well as **phytools** (Revell 2012; Valiente 2010) in order to visualise the network as an interactive tree. Finally, to overlay the tree onto the map, the archived library **R2G2** is used (Arrigo *et al.* 2012; Arrigo 2013). Since haplotypes can be observed in multiple locations, clicking on particular nodes of the tree shows the locations where each copy of the haplotype was found. However, when multiple haplotypes were found in a single location, only one will be displayed, so the `bpec.geoTree` may not tell the whole story. Also note that only existing tip haplotypes are possible to identify on the map.

```
R> bpec.geo <- bpec.geoTree(bpecout, file = "GoogleEarthTree.kml")
```

Tip haplotypes are connected to a tree by a single branch, internal node haplotypes have

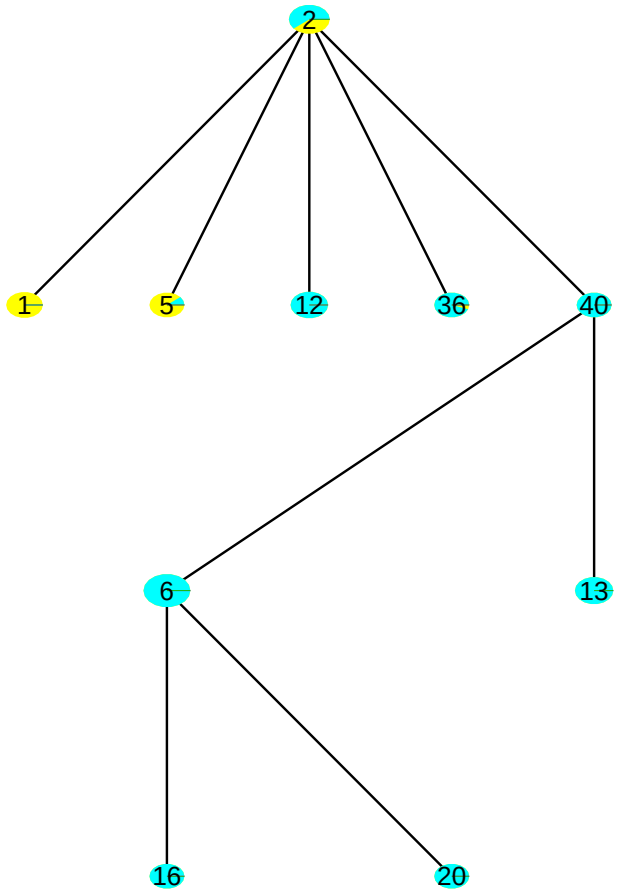

Figure 10: Clustered tree plot of the brown frog dataset using `bpec.treePlot`. Colour corresponds to cluster membership probability and size of node to the number of individuals sampled with each sequence. Edges represent single effective mutations and black dots represent unobserved intermediate haplotypes. In this case all edges have effectively no posterior uncertainty under the model, so they all appear with equal thickness.

three or more connections, whereas branch haplotypes exactly two connections.

6. Analysis of the brown frog data

In the case of the brown frog data, the three locations with highest probability of being ancestral are approximately located at the intersection between the yellow and turquoise cluster, shown as larger dots in Figure 7. These three locations correspond to (a) Paravani lake, treeless mountain steppe, 2100m, Javakheti plateau, posterior probability 24%, (b) Tsalka, treeless mountain steppe, close to the southern slopes of the Trialeti Ridge, posterior probability 12% and (c) Cross Mountain Pass, alpine habitat, 2000-2500m, Great Caucasus, posterior probability 10%. However, it is important to condition any conclusions drawn from these inferences on their associated probability values, and in the case of *R. macrocnemis* it is clear that there is high uncertainty associated with these inferences that is related to the

limited information content of the data rather than issues of convergence. Further sampling (more localities and more individuals per locality) may improve posterior probabilities, but it may also be possible to develop **BPEC** to incorporate outgroup sequences for the inference of ancestral haplotypes (see Section 7), something that should improve posterior probabilities for ancestral areas.

Differences within bioclimatic (annual mean temperature, annual range of temperature, annual precipitation, precipitation seasonality) and altitude variables between the two clusters (Figure 9) are largely due to their mean values rather than their variances. Differences in means are especially apparent for the annual mean temperature (> 5℃ for the yellow cluster and around 5℃ for the turquoise cluster), annual range of temperature (high amplitude of variation, typical for mountain climates: CV $< 320\%$ for the yellow cluster, $> 320\%$ for populations of the turquoise cluster), annual precipitation (higher for populations of the turquoise cluster, nearly 800mm), annual distribution of precipitation (much higher in the yellow cluster, CV $> 45\%$ which results in higher variation in the timing and intensity of annual precipitation). Altitude is rather similar around 1500-1800m. Finally, the landscape dependent variable 'open vs. forested habitat' is clearly different for the clusters. These findings suggest that individuals within the sampled area for *R. macrocnemis* are best described by two geographic clusters of mtDNA sequence variation and that they also differ with respect to specific environmental conditions. These data therefore offer support to the hypothesis that both processes of geographic isolation and divergent selection have contributed to diversification within the group, with the suggestion that taxonomy recognises these entities. As such, the results of **BPEC** provide specific hypotheses than can be further tested with a more extensive genetic marker based approach for hypothesis testing (see Section 7) .

7. Discussion

We have described **BPEC**, an implementation of the phylogeographic and ecological clustering methods described in Manolopoulou *et al.* (2011); Manolopoulou and Emerson (2012). We have introduced several visualization and post-processing tools in order to aid data analysis and interpretation, along with details of the significance of different types of output. **BPEC** will continue to be improved. The main focus of the extensions will revolve around speeding up the convergence of the sampler and improving the approximation stemming from the auxiliary tree parameter.

We recommend caution when extrapolating conclusions from **BPEC** output, and as is the case for many software packages it is important that users do not take a black box approach. Users should condition their conclusions on the biology of their organism of interest, the completeness of their sampling, and the idiosyncrasies of their data (e.g. the proportion of unsampled haplotypes). In terms of extensions to the actual model, more generic evolutionary models for subdivided haplotype trees will be gradually introduced, such as the one recently developed by De Maio *et al.* (2015). Similarly, explicitly modelling the migration process as a spatial transition will allow additional information from the spatial distribution to inform the tree and vice versa. As currently configured, **BPEC** is best treated as a tool that can potentially reduce model space for subsequent hypothesis testing. As an example, in the case of the brown frog *Rana macrocnemis* **BPEC** identified geographic clusters of mtDNA sequence variation that are associated with differing environmental environmental conditions that could underpin divergent selection. Thus, **BPEC** presents evidence that both neutral

and selective processes are driving diversification within the group. However, as **BPEC** is limited to the analysis of a single DNA sequence locus, inferences should not be extrapolated to ultimate biological/ecological conclusions. **BPEC** should lend itself to the integration of inferences across multiple loci within a species, and this is an area that we are investigating for future updates. Of particular relevance is the increasing accessibility of reduced genome sequencing data (McCormack *et al.* 2013) that can provide up to tens of thousands of loci per individual. Filtering for loci characterised by multiple SNPs could provide a rich data source for a multi-locus **BPEC** implementation. Analogous to a single species multi-locus analysis, it should also be possible to integrate across different species sampled from the same locations within **BPEC**. Such an approach would provide for quantitative measures for comparative phylogeography, and this will also be explored for future updates of **BPEC**.

Outgroup sequences can potentially directly inform about the probability of a haplotype being the Most Recent Common Ancestor (MRCA) of a set of sequences, and future versions of **BPEC** will explore the possibility of incorporating outgroup sequences to for this purpose. Sequences immediately derived from an inferred MRCA are also expected to provide some information regarding ancestral areas, and integrating information across the MRCA and sequences immediately derived from it will also be explored.

Finally, we are investigating whether we can extend the applicability of **BPEC** to the analysis of geographic population structure derived from vicariant processes - i.e., where populations become isolated and thus initially share genetic variation, but diverge through time through lineage sorting effects and the accumulation of new population specific mutations. **BPEC** should be applicable to the examination of genealogy among such closely related populations under the evolutionary model of population splitting. In the absence of opposing gene flow among populations, all populations will eventually become diagnosable as descending from a single haplotype unique to that population (lineage sorting). This diagnostic is equivalent to the pattern derived from a colonisation event, and as such it must be borne in mind that clusters defined by **BPEC** may indeed have a vicariant origin. Incorporating a vicariance model into **BPEC** may prove challenging, but it would (i) facilitate the detection of more subtle geographic structuring than that derived from the dispersal model, and (ii) provide a more realistic model of phylogeographic structure.

Acknowledgements

We would like to thank the anonymous reviewers and associate editor for their thoughtful comments which have greatly improved the manuscript. We would also like to thank all the users who have provided feedback on **BPEC** over the years. AH thanks D. Tarkhnishvili for his expert comments on the brown frogs. N. Arrigo, T. Nepusz, L. Revell and G. Valiente quickly answered questions on R functions they published, many thanks to them all.

References

Arrigo N (2013). ***R2G2**: Converting R CRAN Outputs into Google Earth*. R package version 1.0-2, URL https://CRAN.R-project.org/package=R2G2.

Arrigo N, Albert LP, Mickelson PG, Barker MS (2012). "Quantitative Visualization of Bi-

ological Data in Google Earth Using **R2G2**, an R CRAN Package." *Molecular Ecology Resources*, **12**(6), 1177–1179. doi:10.1111/1755-0998.12012.

Atteson K (1999). "The Performance of Neighbor-Joining Methods of Phylogenetic Recronstruction." *Algorithmica*, **25**(2–3), 251–278. doi:10.1007/pl00008277.

Bartholome E, Belward AS (2005). "GLC2000: a new approach to global land cover mapping from Earth observation data." *International Journal of Remote Sensing*, **26**(9), 1959–1977.

Beaumont M (2003). "Estimation of Population Growth or Decline in Genetically Monitored Populations." *Genetics*, **164**(3), 1139–1160.

Bloomquist EW, Lemey P, Suchard MA (2010). "Three Roads Diverged? Routes to Phylogeographic Inference." *Trends in Ecology & Evolution*, **25**(11), 626–632. doi:10.1016/j.tree.2010.08.010.

Boulenger GA (1885). "Description of a new Species of Frog from Asia Minor." In *Proceedings of the Zoological Society of London*, volume 53, pp. 22–23. Wiley Online Library.

Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J (2013). "Hierarchical and Spatially Explicit Clustering of DNA Sequences with **BAPS** Software." *Molecular Biology and Evolution*, **30**(5), 1224–1228. doi:10.1093/molbev/mst028.

Cornuet JM, Pudlo P, Veyssier J, Dehne-Garcia A, Gautier M, Leblois R, Marin JM, Estoup A (2014). "**DIYABC** V2.0: A Software to Make Approximate Bayesian Computation Inferences about Population History Using Single Nucleotide Polymorphism, DNA Sequence and Microsatellite Data." *Bioinformatics*, **30**(8), 1187–1189. doi:10.1093/bioinformatics/btt763.

Csardi G, Nepusz T (2006). "The **igraph** Software Package for Complex Network Research." *InterJournal*, **Complex Systems**, 1695.

De Iorio M, Griffiths R (2004a). "Importance Sampling on Coalescent Histories. I." *Advances in Applied Probability*, **36**(2), 417–433. doi:10.1239/aap/1086957579.

De Iorio M, Griffiths R (2004b). "Importance Sampling on Coalescent Histories. II: Subdivided Population Models." *Advances in Applied Probability*, **36**(2), 434–454. doi:10.1239/aap/1086957580.

De Maio N, Wu CH, O'Reilly KM, Wilson D (2015). "New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation." *PLoS Genetics*, **11**(8), e1005421. doi:10.1371/journal.pgen.1005421.

Desper R, Gascuel O (2002). "Fast and Accurate Phylogeny Reconstruction Algorithms Based on the Minimum-Evolution Principle." In *WABI '02: Proceedings of the Second International Workshop on Algorithms in Bioinformatics*, pp. 357–374. Springer-Verlag.

Drummond AJ, Suchard MA, Xie D, Rambaut A (2012). "Bayesian Phylogenetics with **BEAUti** and the **BEAST** 1.7." *Molecular Biology and Evolution*, **29**(8), 1969–1973. doi:10.1093/molbev/mss075.

Ethier SN, Griffiths RC (1987). "The Infinitely-Many-Sites Model as a Measure-Valued Diffusion." *The Annals of Probability*, **15**(2), 515–545. doi:10.1214/aop/1176992157.

Felsenstein J (1983). "Statistical Inference of Phylogenies." *Journal of the Royal Statistical Society A*, **146**(3), 246–272. doi:10.2307/2981654.

Gascuel O, Steel M (2006). "Neighbor-Joining Revealed." *Molecular Biology and Evolution*, **23**(11), 1997–2000. doi:10.1093/molbev/msl072.

Guillot G, Mortier F, Estoup A (2005). "**GENELAND**: A Computer Package for Landscape Genetics." *Molecular Ecology Notes*, **5**(3), 712–715. doi:10.1111/j.1471-8286.2005.01031.x.

Guindon S, Guo H, Welch D (2016). "Demographic Inference under the Coalescent in a Spatial Continuum." *Theoretical Population Biology*, **111**, 43–50. doi:10.1016/j.tpb.2016.05.002.

Hein J, Schierup M, Wiuf C (2004). *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press.

Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, Rissler L, Victoriano PF, Yoder AD (2010). "Phylogeography's Past, Present, and Future: 10 Years after." *Molecular Phylogenetics and Evolution*, **54**(1), 291–301. doi:10.1016/j.ympev.2009.09.016.

Hijmans RJ (2019). ***raster**: Geographic Data Analysis and Modeling*. R package version 3.0-7, URL https://CRAN.R-project.org/package=raster.

Hijmans RJ, Phillips S, Leathwick J, Elith J (2017). ***dismo**: Species Distribution Modeling*. R package version 1.1-4, URL https://CRAN.R-project.org/package=dismo.

Huelsenbeck J, Ronquist F (2001). "**MrBayes**: Bayesian Inference on Phylogenetic Trees." *Bioinformatics*, **17**(8), 754–755. doi:10.1093/bioinformatics/17.8.754.

Jay F, François O, Durand EY, Blum MGB (2015). "**POPS**: A Software for Prediction of Population Genetic Structure Using Latent Regression Models." *Journal of Statistical Software*, **68**(9), 1–19. doi:10.18637/jss.v068.i09.

Joly S, Stevens MI, van Vuuren BJ (2007). "Haplotype Networks Can Be Misleading in the Presence of Missing Data." *Systematic Biology*, **56**(5), 857–862. doi:10.1080/10635150701633153.

Jombart T, Devillard S, Balloux F (2010). "Discriminant Analysis of Principal Components: A New Method for the Analysis of Genetically Structured Populations." *BMC Genetics*, **11**(1), 94. doi:10.1186/1471-2156-11-94.

Kahle D, Wickham H (2013). "ggmap: Spatial Visualization with ggplot2." *The R Journal*, **5**(1), 144–161. URL https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf.

Kingman J (1982). "The Coalescent." *Stochastic Processes and their Applications*, **13**(3), 235–248. doi:10.1016/0304-4149(82)90011-4.

Knowles LL (2009). "Statistical Phylogeography." *Annual Review of Ecology, Evolution, and Systematics*, **40**, 593–612. doi:10.1146/annurev.ecolsys.38.091206.095702.

Knowles LL, Maddison WP (2002). "Statistical Phylogeography." *Molecular Ecology*, **11**(12), 2623–2635. doi:10.1046/j.1365-294X.2002.01410.x.

Knuth D (1998). "Sorting and Searching." In *The Art of Computer Programming*, volume 3. Addison-Wesley.

Lemey P, Rambaut A, Drummond AJ, Suchard MA (2009). "Bayesian Phylogeography Finds Its Roots." *PLoS Computational Biology*, **5**(9), e1000520. doi:10.1371/journal.pcbi.1000520.

Manolopoulou I, Emerson BC (2012). "Phylogeographic Ancestral Inference Using the Coalescent Model on Haplotype Trees." *Journal of Computational Biology*, **19**(6), 745–755. doi:10.1089/cmb.2012.0038.

Manolopoulou I, Legarreta L, Emerson BC, Brooks S, Tavaré S (2011). "A Bayesian Approach to Phylogeographic Clustering." *Interface Focus*, **1**(6), 909–921. doi:10.1098/rsfs.2011.0054.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT (2013). "Applications of Next-Generation Sequencing to Phylogeography and Phylogenetics." *Molecular Phylogenetics and Evolution*, **66**(2), 526–538. doi:10.1016/j.ympev.2011.12.007.

Papastamoulis P, Iliopoulos G (2010). "An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions." *Journal of Computational and Graphical Statistics*, **19**(2), 313–331. doi:10.1198/jcgs.2010.09008.

Plummer M, Best N, Cowles K, Vines K (2006). "**coda**: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11.

Pritchard JK, Stephens M, Donnelly P (2000). "Inference of Population Structure Using Multilocus Genotype Data." *Genetics*, **155**(2), 945–959.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Revell LJ (2012). "**phytools**: An R Package for Phylogenetic Comparative Biology (and Other Things)." *Methods in Ecology and Evolution*, **3**(2), 217–223. doi:10.1111/j.2041-210x.2011.00169.x.

Rzhetsky A, Nei M (1993). "Theoretical Foundation of the Minimum-Evolution Method of Phylogenetic Inference." *Molecular Biology and Evolution*, **10**(5), 1073–1095. doi:10.1093/oxfordjournals.molbev.a040056.

Stephens M (2000a). "Bayesian Analysis of Mixture Models with an Unknown Number of Components – An Alternative to Reversible Jump Methods." *The Annals of Statistics*, **28**(1), 40–74. doi:10.1214/aos/1016120364.

Stephens M (2000b). "Dealing With Label-Switching in Mixture Models." *Journal of the Royal Statistical Society B*, **62**(4), 795–809. doi:10.1111/1467-9868.00265.

Tarkhnishvili D, Hille A, Böhme W (2001). "Humid Forest Refugia, Speciation and Secondary Introgression Between Evolutionary Lineages: Differentiation in a Near Eastern Brown Frog, *Rana Macrocnemis*." *Biological Journal of the Linnean Society*, **74**(2), 141–156. doi:10.1111/j.1095-8312.2001.tb01383.x.

Valiente G (2010). *Combinatorial Pattern Matching Algorithms in Computational Biology Using Perl and R*. CRC Press.

Veith M, Kosuch J, Vences M (2003a). "Climatic Oscillations Triggered Post-Messinian Speciation of Western Palearctic Brown Frogs (Amphibia, Ranidae)." *Molecular Phylogenetics and Evolution*, **26**(2), 310–327. doi:10.1016/s1055-7903(02)00324-x.

Veith M, Schmidtler J, Kosuch J, Baran I, Seitz A (2003b). "Palaeoclimatic Changes Explain Anatolian Mountain Frog Evolution: A Test for Alternating Vicariance and Dispersal Events." *Molecular Ecology*, **12**(1), 185–199. doi:10.1046/j.1365-294x.2003.01714.x.

Wakeley J (2013). "Coalescent Theory Has Many New Branches." *Theoretical Population Biology*, **87**, 1–4. doi:10.1016/j.tpb.2013.06.001.

8. Appendix

8.1. Temporal orderings

Suppose the haplotype tree is given by the top tree of Figure 11 (Manolopoulou and Emerson 2012). For ease of exposition, the numbers on the nodes here represent the sample sizes of each haplotype rather than the label of each haplotype.

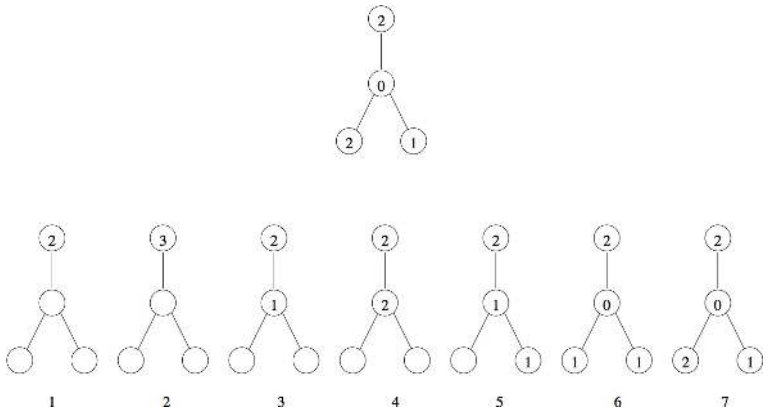

Figure 11: Top panel: In this tree the oldest haplotype of the sample (the top haplotype) is observed twice in the sample, whereas the intermediate haplotypes in not observed at all. Bottom panel: a possible time evolution of how the haplotypes arose. Nodes without a number correspond to haplotypes which haven't appeared yet. At first one sequence is present, the ancestral sequence, which replicated into two (the first event is always a replication, otherwise that haplotype would disappear). Then one of those two identical sequences may replicate again to give us a total of three (or could have mutated to give a new haplotype). One of those three then mutate to give us the intermediate haplotype, which in turn here replicates and then mutates (and goes extinct) to give us the right-hand leaf. Finally, the intermediate haplotype mutates again to give us the left-hand leaf, which then also replicates to give another copy of itself.

Simulating a temporal ordering implies that, starting with the ancestral sequence, we specify a series of replication and mutation events which occurred by mimicking evolution, eventually resulting in the observed haplotype tree. A possible series of events is shown in Figure 11 through 7 timepoints, where node numbers indicate number of copies of each haplotype.

Notice that, if the root node had replicated further, we would have had three copies of the root haplotype. Although in theory this could have happened, with one of the copies eventually becoming extinct, we do not take into account any such scenarios, instead we only account for the observed sequences. Additionally, it would not have been possible for the intermediate haplotype to mutate after Step 3 above, since then it would disappear from the ancestral sequences, and another mutation would not have been possible.

8.2. Computational issues

Haplotype tree likelihood

When calculating the Metropolis-Hastings ratio for a proposal from root r to r', one needs to calculate

$$\frac{p(r' \mid T, \mathcal{S})}{p(r \mid T, \mathcal{S})} = \frac{\frac{\left|\mathcal{O}^{\mathcal{S}}_{r',T}\right|}{\sum_r \left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}}{\frac{\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}{\sum_r \left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}} = \frac{\left|\mathcal{O}^{\mathcal{S}}_{r',T}\right|}{\left|\mathcal{O}^{\mathcal{S}}_{r,T}\right|}.$$

However, computing the size of the two sets of temporal orderings is a computational bottleneck. To overcome this issue, an unbiased estimator of the likelihood is used instead. Conditionally on a root r and tree T, a particular ordering O^* is generated by moving from the root to the tips and randomly choosing among the available replicate/mutate moves at each step, according to some distribution $q(O^*)$, such that any possible ordering of $\mathcal{O}^{\mathcal{S}}_{r,T}$ can arise. Then calculate

$$\widehat{\left|\mathcal{O}^{\mathcal{S}}_{r',T}\right|} = \frac{1}{q(O^*)},$$

such that

$$\mathbb{E}\left(\widehat{\left|\mathcal{O}^{\mathcal{S}}_{r',T}\right|}\right) = \sum_{O^* \in \mathcal{O}^{\mathcal{S}}_{r,T}} \frac{1}{q(O^*)} \times q(O^*) = \left|\mathcal{O}^{\mathcal{S}}_{r',T}\right|,$$

so it provides an unbiased estimator of the likelihood. Note here that the latent variable O is not accepted/rejected together with the root, so the Markov chain Monte Carlo does not maintain detailed balance (see Manolopoulou and Emerson 2012; Beaumont 2003). This is because variance of $q(O^*)$ can be huge and detrimental to the MCMC, causing it to get 'stuck'; future improvements of $q(O^*)$ could allow O^* to be accepted/rejected together with the root r. Multiple realizations of O^* could also be used instead, but **BPEC** only considers 1.

MCMC exploration and convergence

The **BPEC** model faces two additional key computational bottlenecks. The first comes from learning the posterior probability of the root haplotype. Since it relies upon an estimator of the likelihood, a large number of iterations are required in order to allow for reasonable convergence. However, the total number of haplotypes (and as such the number of possible roots) is generally low (usually up to a few hundreds), so with enough iterations the sampler can explore the whole root parameter space sufficiently.

On the other hand, the clustering parameter space is challenging to adequately explore. Instead, sophisticated local proposals are required. Manolopoulou *et al.* (2011) implement a clustering proposal which cumulatively adds observation branches (as shown in Figure 5) to clusters by starting with empty clusters with mean and variance equal to their corresponding prior means. As each observation branch is added to one of the clusters (in random order), the means and variances of that cluster are updated according to the corresponding posterior means. This allows the sampler to propose clusters for each branch according to the cluster in which it fits best, while randomising the order of the allocation meant that no

branches were given higher weight than others. In **BPEC** we tweak the proposal distribution of Manolopoulou *et al.* (2011) by introducing an auxiliary variable w_c, representing the weight of the previous clustering in the MCMC sampler. Rather than simply allocating each observation branch to one of the existing clusters simply by assessing the fit of each branch to each of the clusters, we assign it to the same cluster as the previous iteration (where possible) with probability w_c. This favours clusterings similar to the previous iteration, thereby ensuring that local moves are proposed more frequently. Since w_c is an auxiliary variable, it is accepted/rejected together with the proposed parameters, so the sampler automatically chooses a value of w_c that is reasonable.

Label-switching

In order to draw cluster-specific inferences, cluster labels need to be assigned for every posterior sample available. This is known as the label-switching problem (Stephens 2000a,b; Papastamoulis and Iliopoulos 2010) and it is especially challenging in the case of a variable number of clusters. Here we take a pivoting approach to assign cluster labels on-line (i.e., without the need of post-processing). The algorithm works as follows:

1. During burn-in of the first chain, record the cluster labels of the posterior sample with the highest value of the posterior density, denoted by $\mathbf{c}^*$.

2. Once this clustering is fixed, subsequent labels of the posterior sample of the set $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are chosen such that
$$p\left(\mathbf{Y} \mid \boldsymbol{\mu}_{c^*}, \boldsymbol{\Sigma}_{c^*}, c^*\right)$$
is maximised. In other words, labels of the set of means and covariances are chosen such that the likelihood relative to (approximate) maximum a posteriori clustering c^* is maximised.

Hashing

In contrast to coalescent trees, which are binary and can be represented simply by the pairs of subsequent coalescence events, haplotype trees do not have shorthand representations. Instead, a standard way to represent a haplotype tree is through its corresponding graph adjacency matrix. However, keeping track of posterior samples of trees requires storing entire matrices at each iteration of the sampler, which creates a memory bottleneck.

In our case, we can take advantage of the fact that not all adjacency matrices are possible; most edges are either certainly present or absent as determined by Ω. Uncertainty only arises through edges that are part of a loop in the network, so each tree is characterised by the set of deleted edges. Trees are then reduced to vectors of length n_{loop} with integer entries. Standard hashing techniques can thus be used to store the number of times each tree (i.e., each integer vector) appears in the MCMC posterior samples.

Hashing algorithms allow us to represent integer vectors by an single integer. In our case, we can store the index of the edge deleted from each loop at each iteration of the MCMC, keeping track of them via the 'hashing index' of the entire vector. Hash functions create a short (as short as possible) address book where each of these numbers is stored in a specific page, in such a way that it can easily be retrieved (see Knuth 1998).

Affiliation:

Ioanna Manolopoulou
Department of Statistical Science University College London
London WC1E 6BT, UK
E-mail: ioanna@stats.ucl.ac.uk
URL: http://www.ucl.ac.uk/~ucakima/

Axel Hille
Institute of Applied Statistics Dr. Jörg Schnitker Ltd.
Oberntorwall 16-18
D-33602 Bielefeld, Germany
E-mail: axel.hille@gmx.net

Brent C. Emerson
Island Ecology and Evolution Research Group
Instituto de Productos Naturales y Agrobiologia
C/Astrofisico Francisco Sanchez 3
La Laguna
Tenerife
Canary Islands 38206
Spain, and
School of Biological Sciences
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
UK
E-mail:bemerson@ipna.csic.es