

# Phylogeographic Ancestral Inference Using the Coalescent Model on Haplotype Trees

IOANNA MANOLOPOULOU and BRENT C. EMERSON

## ABSTRACT

Phylogeographic ancestral inference is an issue frequently arising in population ecology that aims to understand the geographical roots and structure of species. Here, we specifically address relatively small scale mtDNA datasets (typically less than 500 sequences with fewer than 1000 nucleotides), focusing on ancestral location inference. Our approach uses a coalescent modelling framework projected onto haplotype trees in order to reduce computational complexity, at the same time adhering to complex evolutionary processes. Statistical innovations of the last few years have allowed for computationally feasible yet accurate inferences in phylogenetic frameworks. We implement our methods on a set of synthetic datasets and show how, despite high uncertainty in terms of identifying the root haplotype, estimation of the ancestral location naturally encompasses lower uncertainty, allowing us to pinpoint the Maximum A Posteriori estimates for ancestral locations. We exemplify our methods on a set of synthetic datasets and then combine our inference methods with the phylogeographic clustering approach presented in Manolopoulou et al. (2011) on a real dataset from weevils in the Iberian peninsula in order to infer ancestral locations as well as population substructure.

**Key words:** phylogeography, ancestral inference, statistics, coalescent.

## 1. INTRODUCTION

PHYLOGEOGRAPHIC ANCESTRAL INFERENCE IS A QUESTION frequently arising in population ecology that is concerned with the geographical roots and structure of species. While many of the cutting-edge approaches (Bloomquist et al., 2010; Lemey et al., 2010, 2009) can provide powerful inferences using sophisticated evolutionary models, the amount of data (especially when sampling resources are limited) naturally leads to high levels of uncertainty. At the same time, although parsimonious approaches (Swofford and Berlocher, 1987), typically represented by gene trees, are computationally efficient, it is well-known that they can misrepresent evolution, leading to biased inferences (Felsenstein, 1978). Here, we specifically address relatively small scale mtDNA datasets (typically less than 500 sequences with fewer than 1000 nucleotides), focusing on ancestral location inference.

Our approach uses a coalescent modelling framework projected onto haplotype trees in order to reduce computational complexity, but adhering to complex evolutionary processes. Statistical innovations of the

---

Department of Statistics, Duke University, Durham, North Carolina.

last few years have allowed for computationally feasible yet accurate inferences in phylogenetic frameworks. Although our approach does not employ a rigorous geographical migration model, analysis of synthetic datasets shows that our methods can provide valid and computationally efficient results. Despite high uncertainty in terms of identifying the root node, estimation of the ancestral location naturally encompasses lower uncertainty, allowing us to pinpoint the Maximum A Posteriori estimates for ancestral locations. We combine our inferences with the phylogeographic clustering methods presented in Manolopoulou et al. (2011) on a real dataset from weevils in the Iberian peninsula in order to infer ancestral locations as well as population substructure.

The article is organized in the following sections: Section 1 presents the core of the evolutionary model used and its projection onto haplotype trees, and Section 2 proceeds to incorporate ancestral location inference. Section 3 presents analysis of a set of synthetic datasets, which is followed by the analysis of a real dataset in Section 4.

## 2. HAPLOTYPE TREE MODEL

Coalescent theory (Kingman, 1982), in its basic form, retrospectively relates a set of sequences back to their Most Recent Common Ancestor (MRCA) through a series of coalescence events, based on the assumptions of constant population and random mating. In a sample of  $N$  sequences viewed backwards into the past, the length of time to the latest coalescence event is proportional to  $\binom{N}{2}$ . The limited size of the datasets at hand imply that multiple sampled individuals may correspond to the same haplotype, so that some individuals may only be identifiable in terms of their sampling location. In order to use haplotype trees as a representation of evolution faithful to the coalescent model (Wakeley, 2008; Kingman, 1982), we describe a theoretical framework whereby haplotype trees are translated into sets of coalescent trees.

Assuming that mutations occur independently as a Poisson process at rate  $\theta/2$ , they can be thought of as being poured down the coalescent tree (Tavaré, 1986, 2003) resulting in a combined Markov process involving both coalescence and mutation events. A rooted haplotype tree only partially determines a set of mutation and coalescence events which occurred in history, and places restrictions on their ordering (see Appendix A). Calculating probabilities over rooted haplotype trees therefore requires summing over all possibilities and orderings of past events. We denote the temporal order of mutation and split events with  $\mathcal{H}$ , with  $\mathcal{H}^t$  being the  $t^{\text{th}}$  event and  $H$  the total number of events. We use a simple evolutionary model with equal exponential mutation process rates across all nucleotide sites and between all possible mutations, but the flexibility of our methods easily extends to more complex evolutionary processes. Conditional on the root  $r$  and the total number of events  $H$  we have

$$\mathbb{P}(\mathcal{H}|H, r) = \prod_t \mathbb{P}(\mathcal{H}^t|r), \quad (1)$$

where the events, forwards in time (Ethier and Griffiths, 1987; Stephens and Donnelly, 2000; Tavaré, 2003), are independent and have probabilities given by the ratios of the underlying exponential rates  $\binom{N_t}{2}$  and  $\frac{\theta N_t}{2}$  ( $N_t$  being the number of sequences present at time  $t$ ), so that

$$\begin{aligned} \mathbb{P}(\text{a given sequence splits}) &= \frac{1}{N_t} \frac{N_t - 1}{(N_t - 1 + \theta)}, \\ \mathbb{P}(\text{a given sequence mutates at any of its sites}) &= \frac{1}{N_t} \frac{\theta}{(N_t - 1 + \theta)}. \end{aligned} \quad (2)$$

This implies that the probability of a haplotype tree, denoted by  $T$ , given the root  $r$ , can be calculated by summing over the probabilities (1) of all temporal orderings  $\mathcal{H}_j$  which are consistent with the tree. In other words,

$$\mathbb{P}(T|r) = \sum_j \mathbb{P}(\mathcal{H}_j|H_j, r), \quad (3)$$

where  $\mathcal{H}_j$  is consistent with  $T$ . This model implicitly incorporates information about the number of copies of each haplotype within the sample.

We begin by assuming that, in the absence of any information about the mutation process, any haplotype tree  $T$  is equiprobable, and place an Inverse-Gamma prior on the mutation rate  $\theta \sim IG(a_\theta, b_\theta)$ . Similarly to the haplotype tree, we assume that all tree topologies are equally likely a priori given the root, so that

$p(T|r) \propto 1$ , and any sequence (as opposed to haplotype) is equally likely to be the root, so that  $p(r) \propto 1$ . The distribution of the sequence data  $\mathcal{S}$  becomes

$$P(\mathcal{S}|T, r) = \begin{cases} 1 & \text{if } T \text{ consistent with } \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

Although only one sequence set  $\mathcal{S}$  is consistent with the fully specified haplotype tree  $T$ , there are several haplotype trees represented by the set  $\Omega$  consistent with the sequences. The posterior distribution of the haplotype tree then becomes

$$P(T|\mathcal{S}, r) \propto \begin{cases} \mathbb{P}(T|r) & \text{if } T \in \Omega \\ 0 & \text{otherwise,} \end{cases}$$

where the normalization constant may be calculated as  $\sum_{T_i \in \Omega} \mathbb{P}(T_i|r)$ . Similarly, we can calculate the posterior distribution for the root

$$P(r|\mathcal{S}, T) \propto P(T|r).$$

Calculation and exploration of the infinite state space  $\Omega$  which is consistent with  $\mathcal{S}$  is computationally challenging. We contend that, under an argument of relaxed parsimony, it is possible to reduce the state space to a finite (but vast) set  $\Omega := \Omega(\mathcal{S})$  of realistic haplotype trees (Manolopoulou et al., 2011).

Finally, the probability of a tree can only be calculated conditional on the total number of mutation and split events, here denoted by  $H$ . However, the set  $\Omega$  may contain trees involving a different number of events. In order to calculate the probability of any one of those trees, we require

$$\mathbb{P}(T|r) \propto \mathbb{P}(T|H, r) \times \mathbb{P}(H|r).$$

Assuming a uniform prior on  $H$  such that

$$\mathbb{P}(H|r) \propto 1,$$

we see that  $\mathbb{P}(T|r)$  can be calculated using (3) and simply multiplying over all the events for any size of tree  $H$ . Although this prior distribution is improper since it covers an infinite number of trees, the posterior distribution is guaranteed to be proper because the set  $\Omega := \Omega(\mathcal{S})$  of haplotype trees is finite.

This model provides several contributions to inference on the rooted haplotype tree. Based on explicit distributions about haplotype trees, it supplies a rigorous mathematical framework for estimation. It is consistent with many of the theoretical properties of ancestral inference developed by Griffiths and Tavaré (1994), and empirical predictions on haplotype trees raised by Crandall and Templeton (1993) and Posada and Crandall (2001). For example, older alleles have a greater probability of becoming interior (as opposed to leaf) haplotypes: this may be directly derived from (3), since interior haplotypes naturally allow a much larger number of orderings in which events may have occurred. Furthermore, haplotypes of greater frequency are more likely to have a higher degree (i.e., more mutational connections in the tree): the probability of a mutation increases according to frequency of the haplotype. Perhaps the most important advantage of this model is that the posterior probability of a haplotype tree (1) can be explicitly expressed, which allows for backward rather than forward inference. Finally, although the estimates for the root are inherently unreliable because of the variation in the model, the synthetic data analysis in Section 3 shows that when the sequence data  $\mathcal{S}$  are combined with geographical data for each individual, ancestral locations may be estimated with a much higher probability of success.

In order to draw inferences about the haplotype tree under this model, calculation of  $\mathbb{P}(T|\mathcal{S}, r)$  is required. We employ an auxiliary variable approach following Beaumont (2003), whereby a single temporal ordering of haplotype tree events is considered an auxiliary variable in our inferences, allowing for computational feasibility. An improved approach would allow a set of several temporal orderings (Beaumont, 2003) to be considered as latent in order to reduce Monte Carlo error; previous analyses (Manolopoulou, 2009) showed that, in our case, this was not necessary. Our framework then fully specifies the distribution

$$p(r, \mathcal{H}, T|\mathcal{S}).$$

In order to explore the space of temporal orderings of events, we describe Algorithm 1.1, similar to Ethier and Griffiths (1987), which generates temporal orderings consistent with a fixed haplotype tree  $T$ .

---

**Algorithm 1.1. Random Temporal Ordering**


---

This algorithm generates temporal orderings by mimicking the ancestral history of the sample, starting with the root and ending with the observed sequences.

1. Start at the root. Initially only one copy of the root haplotype is present. Split it into two copies and repeat the next step until all mutation or split events determined by the haplotype tree have occurred.
  2. For all sequences present, consider all mutations and splits that are consistent with the haplotype tree. Draw one of those events proportional to the probabilities given by (2), and repeat this step until the tree is complete.
- 

For each temporal ordering  $\mathcal{H}$  generated from this algorithm, the proposal probability  $q(\mathcal{H}|T, r)$  can be calculated by multiplying the normalized terms in (2) over all events. By construction, any  $\mathcal{H}$  consistent with  $T$  may be generated under Algorithm 1.1 because at all the steps, the consistent events have non-zero probability.

In order to simulate samples from the posterior distribution of  $(r, T)$ , we run a Markov chain Monte Carlo sampler.

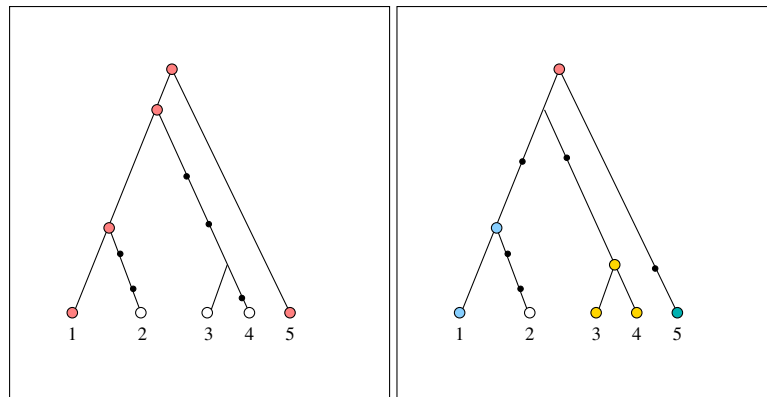
### 3. ANCESTRAL LOCATIONS IN PHYLOGEOGRAPHIC ANALYSIS

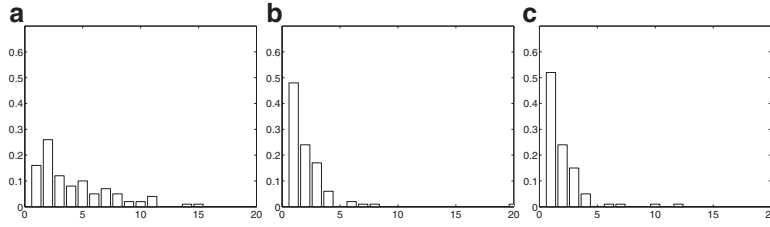
One of the objectives of phylogeographic ancestral analysis is to identify the location(s) where a population originated. However, datasets at hand do not show enough structure in order to implement subdivided population and migration models such as found in Bahlo and Griffiths (2000) and De Iorio and Griffiths (2004), especially with small sample sizes and geographically neighboring locations. The analysis presented here does not assume a geographical model for the spread of populations in time, but allows for heuristically inferring ancestral locations (in terms of specific sampling points) by tracing the locations of the root haplotype, or, if extinct, the locations of its following descendants.

In cases where the root haplotype is not missing, as in the left-hand panel of Figure 1, we know that the ancestral location is more likely to contain that haplotype. For example, if a location contains both copies of the pink haplotype, it is more likely to be ancestral than a location which only contains younger haplotypes. If, on the other hand, the root haplotype is missing, as in the right-hand panel of Figure 1, then the ancestral location is more likely to contain immediate descendants. This approach is consistent with many descriptive characteristics of an ancestral area (Emerson and Hewitt, 2005).

Our heuristic approach calculates the contribution of each of the oldest observed haplotypes along all possible branches originating at the root (if the root haplotype is observed in our sample, then we simply have the root only), and then adds the contribution of each of those haplotypes for each location. For example, referring back to the right-hand panel of Figure 1, and assuming that the light blue and one of the yellow sequences appeared in location A, green in location B, and remaining yellow in location C, then the

**FIG. 1.** Two possible genealogy scenarios, where terminal nodes represent observed sequences (with the color representing the haplotype), and small black circles are mutations. In the figure on the left, the oldest haplotype is the pink one at the top, which appears twice in the sample. In the figure on the right, the oldest haplotype is missing, and the next possible descendants are the blue, yellow, and green haplotypes.





**FIG. 2.** (a) Histogram of Maximum A Posteriori (MAP) estimates for the root haplotype in the case of a known tree. (b, c) MAP estimates of ancestral locations for a known and unknown haplotype tree. Note that assessment of the ancestral haplotype inference in the case of an unknown tree is not possible, as haplotypes cannot always be uniquely identified.

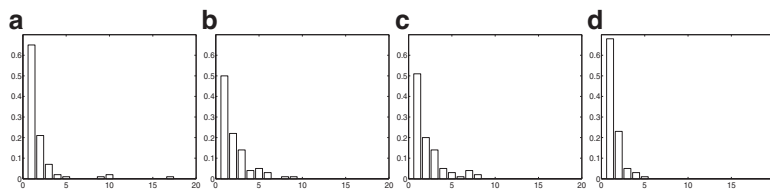
contribution of each of those haplotypes to each of the locations would be  $1/4$ , so that the three probabilities for locations A, B, and C become 0.5, 0.25, and 0.25, respectively. If, on the other hand, the light blue sequence appeared in location A, the 2 yellow ones in location B, and green in location C, the three probabilities for locations A, B, and C become 0.33, 0.33, and 0.33, respectively. Although we do not take into account distance from the root, geographical location, or number of times each haplotype appears in a location, implicitly assuming standing variation in the population, our synthetic trials have shown that our approach provides valuable results in inferring ancestral locations.

#### 4. SYNTHETIC DATA ANALYSIS

We generate a set of 100 replicate synthetic datasets and assess the performance of our algorithm. Each dataset is initiated by a sequence of length  $l = 500$ , at an initial geographical location  $y_{11} = (0, 0)$  and with mutation rate  $\frac{\theta}{2} = 1.8$ . Each new sequence  $j$  of haplotype  $i$  then is assumed either to stay in its current location, or move to a new location: with probability 0.85 it stays in the geographical location of its ancestor  $a_{ij}$  such that  $y_{ij} = y_{a_{ij}}$  (implying a migration rate of 0.15); otherwise, it moves to a new location  $y_{ij} = N(y_{a_{ij}}, 0.1)$ . The new sequence is forced to start a new location if the location of its ancestor contains 15 or more sequences. Extinction occurs when all of the branches of the gene tree carrying the same haplotype undergo a mutation. These tuning generative parameters were chosen in order for the synthetic datasets to match the real dataset at hand as much as possible. The iterative algorithm stops when it reaches 100 observed sequences (not including ones which are extinct in the process), corresponding to a variable number of haplotypes, locations and geographical clusters. Locations and haplotypes are ordered from oldest to most recent.

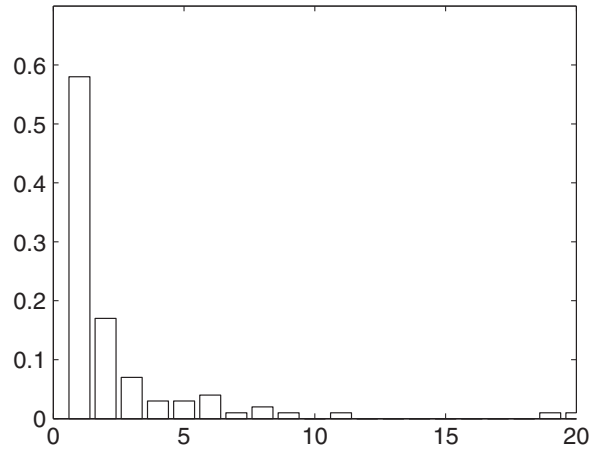
We assume both a known and an unknown tree, showing the results below. As expected, the lack of sufficient data results in weak performance at inferring root haplotypes, as indicated by Figure 2a showing Maximum A Posteriori (MAP) estimates for the root haplotype. However, the methods are successful in inferring ancestral locations. In the case of a known haplotype tree shown in Figure 2b, the top three MAP ancestral locations cover over 90% of the cases and the corresponding percentage in the case of an unknown haplotype tree is roughly 85% (Fig. 2c). Although the data generative algorithm used here results in much stronger phylogeographic structure than observed in real datasets, it is nevertheless important to verify that our methods are able to successfully identify ancestral locations without the use of computationally expensive models.

We now depart from this initial set of parameters and investigate the effect of mutation and migration rates. Specifically, we set the mutation rate to  $\theta/2 = 1.2$  and  $\theta/2 = 3$  keeping all other parameters constant, with results shown in Figure 3a,b. As expected, higher mutation rates lead to higher haplotype variability and thus higher uncertainty in ancestral locations. We then fix  $\theta/2 = 1.8$  and use migration rates 0.8 and 0.9



**FIG. 3.** (a, b) Maximum A Posteriori histograms of the top 20 ancestral haplotypes with  $\theta/2 = 1.2$  and  $\theta/2 = 3$  respectively, migration rate 0.85. (c, d) show corresponding histograms for  $\theta/2 = 1.8$  and migration rates 0.8 and 0.9, respectively.

**FIG. 4.** Histograms of Maximum A Posteriori estimates of the top 20 ancestral haplotypes using the GTR mutation model with total migration rate  $\theta/2 = 1.8$  and migration rate 0.85.



respectively, with corresponding histograms shown in Figure 3c,d, revealing the intuitive fact that higher migration rates lead to higher ancestral location uncertainty. Note that, although the range of values for the mutation and migration rates is relatively narrow, it reflects a wide range of scenarios due to the sensitivity of the model on those parameters.

Finally, we use a Generalized Time-Reversible (GTR) mutation model (Tavaré, 1986) in order to address possible biases introduced by unequal mutation rates. Mutations occur as a Markov Process with generator Q-matrix

$$Q = \phi_j \begin{pmatrix} \cdot & v_1 \pi_G & v_2 \pi_C & v_3 \pi_T \\ v_1 \pi_A & \cdot & v_4 \pi_C & v_5 \pi_T \\ v_2 \pi_A & v_4 \pi_G & \cdot & v_6 \pi_T \\ v_3 \pi_A & v_5 \pi_G & v_6 \pi_C & \cdot \end{pmatrix}$$

where the  $\pi_i$  ( $i = A, G, C, T$ ) represent the equilibrium probabilities of the nucleotides, the mutation coefficients  $v_1, \dots, v_6$  the relative mutation probabilities and  $\phi_j$  site-specific mutation coefficients. More information on the GTR mutation model can be found in Appendix C. We use the following distributions for the unknown parameters

$$\begin{aligned} (\pi_1, \pi_2, \pi_3, \pi_4) &\sim \mathcal{D}(2, 2, 2, 2) \\ v_1, v_6 &\sim \mathcal{G}(1, 1) \\ v_2, v_3, v_4, v_5 &\sim \mathcal{G}(1, 7) \\ \phi_I &\sim \mathcal{G}(1, 0.2), \end{aligned}$$

so that the total migration rate  $\theta/2 \approx 1.8$ , but unevenly distributed and accounting for transition bias. Here  $\mathcal{G}$  and  $\mathcal{D}$  represent the Gamma and Dirichlet distributions respectively. The corresponding histogram of ancestral locations is shown in Figure 4. Despite the unequal mutation rates, results are very similar to the case of equal mutation rates. This is possibly due to the time span and size of dataset being too limited for the variable mutation rates to have a significant impact on the simulated datasets.

## 5. REAL DATASET IMPLEMENTATION

We apply our algorithms to a mitochondrial DNA dataset of weevils in the Iberian peninsula. *Rhinusa vestita* is a seed parasite weevil feeding and reproducing on snapdragons. It is believed to have been present in Portugal, Spain, France, and Italy. The complete nucleotide sequence for the mitochondrial COII gene (722 bp) was obtained for 275 *Rhinusa vestita* individuals. Previous studies investigating the association of weevils with three host plant species, combined with knowledge about the glaciation history of the Iberian peninsula (Hewitt, 2000), led to the biological prediction that the species originated from the Rhône valley to the east and west.

TABLE 1. POSTERIOR ANCESTRAL PROBABILITIES OF THE TOP FOUR SAMPLING LOCATIONS OF THE *RHINUSA VESTITA* DATA

<i>Location</i>	<i>Posterior mass</i>
Brissac	0.26
Petit Luberon	0.21
La Clape	0.14
Grotte Petit	0.14

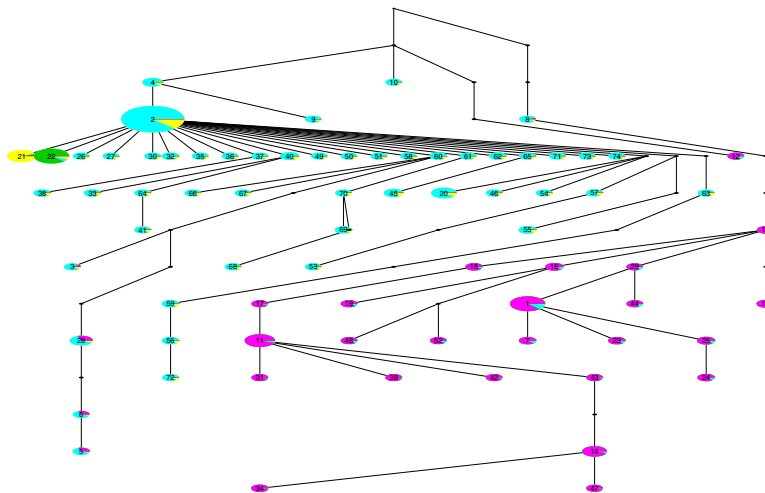
We combine our methods with the analysis presented in Manolopoulou et al. (2011), in order to infer both ancestral locations but also population substructure. Manolopoulou et al. (2011) describe a model for inferring phylogeographic substructure by identifying population clusters consistent with individual migrations, but without providing temporal results such as ancestral locations. Here, we combine these methods with our ancestral inference within the same Markov chain Monte Carlo inference, in order to provide results both on population substructure as well as geographical origin.

The results confirm the biological hypothesis of the location of origin; the top four sampling locations, collecting 75% of posterior mass, are shown in Table 1. As in Manolopoulou et al. (2011), we plot what is now a rooted gene tree shown in Figure 5, using color to indicate geographical population cluster (in contrast to Fig. 1, where color represented haplotype). Here, each color represents a geographically relatively homogeneous population, and each new cluster (and corresponding color) is formed via migration. Figure 6 shows the corresponding geographical distribution of each of the population clusters, with the larger spot representing the most likely ancestral location. Combined with our ancestral location inference methods, it appears that weevils originated around the Rhône valley, migrating to the east and west, in agreement with biological predictions.

## 6. DISCUSSION

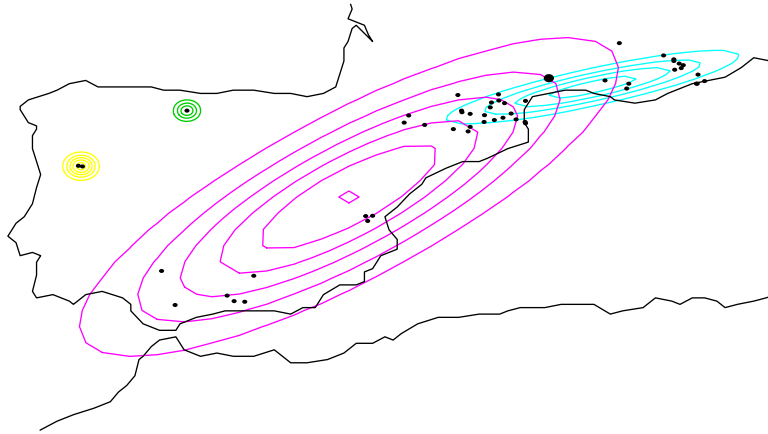
We have presented a statistical framework whereby the coalescent model is used in order to draw inferences about haplotype trees through Markov chain Monte Carlo. In addition, we have described methods for inferring ancestral locations in phylogeographic settings. Our results were validated by simulated synthetic datasets, and were successful in confirming the biological hypothesis in the real dataset.

Although more sophisticated evolutionary models may be used to account for a variable population size (Slatkin, 2001), selection (Neuhausser and Krone, 1997), and recombination (Hudson and Kaplan, 1988), prior implementations we ran showed that in small-scale datasets such as the one at hand, the data are very weakly informative about many of the additional evolutionary parameters. Perhaps the most valuable extension would allow for the coalescence rate to vary across population clusters, in order to represent local



**FIG. 5.** One of the non-unique MAP estimates of the haplotype tree using our approach, where color corresponds to cluster and size to the number of individuals sampled with each sequence.

**FIG. 6.** Corresponding bivariate normal contour plots evaluated at the posterior means for the weevil data-set. The black dots indicate sampling locations, and colors correspond to the clusters shown in Figure 5. The largest dot corresponds to the MAP ancestral location.



proliferations. Similarly, rigorous theoretical calculations relating ancestral haplotypes with geographical locations (Bloomquist et al., 2010; Lemey et al., 2010, 2009), perhaps through the use of an explicit migration model, can provide a solid basis for an improved estimator of ancestral locations.

Finally, our methods are freely available through an R package “Bayesian Phylogeographic Clustering” (available at [www.stat.duke.edu/~im30/software.html](http://www.stat.duke.edu/~im30/software.html)).

## APPENDIX

### A. Haplotype tree example

Suppose the haplotype tree is given by the top tree of Figure 7. For ease of exposition, the numbers on the nodes here represent the sample sizes of each haplotype rather than the label of each haplotype, and we represent each event by updating the numbers on each haplotype according to the number of times it is observed at each time-point in the sample.

Simulating a temporal ordering implies that, starting with the ancestral sequence, we specify a series of split and mutation events which occurred by mimicking evolution, eventually resulting in the fixed haplotype tree. For example, the bottom panel of Figure 7 is a possible temporal ordering of the observed tree given in the top panel.

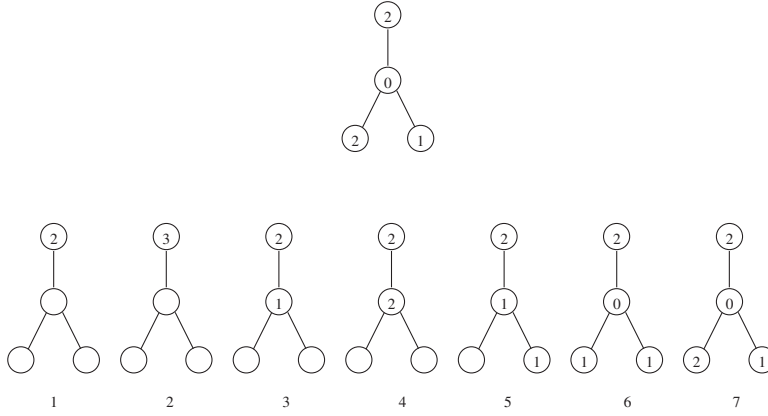
Observe now that, for example, the root node could not have split any further: this would result in three copies of the ancestral haplotype, which is inconsistent with the haplotype tree which specifies precisely two. In addition, it would not have been possible for the intermediate haplotype to mutate after Step 3 above, since then it would disappear from the ancestral sequences, and another mutation would not have been possible. In other words, consistent events are defined as follows.

- A split event is consistent with the haplotype tree, if it does not imply that the sample size of that haplotype will exceed the number of times it appears in the complete haplotype tree, plus the number of mutations that haplotype will be forced to undergo in following steps (so, in the example, the intermediate haplotype after Step 5 will be forced to undergo exactly one more mutation).
- Similarly, a mutation is possible if (a) is true, and (b) OR (c) are true:
  - (a) it is represented by an edge on the haplotype tree, where the ancestral sequence of the edge has already appeared in the ancestral sample;
  - (b) the ancestral sequence of the edge corresponding to that mutation does not go extinct;
  - (c) the ancestral sequence of the edge goes extinct, and there are not more events involving that sequence which have not yet occurred but are forced by the haplotype tree.

### B. Markov chain Monte Carlo sampler

The complete model contains the tree topology  $T$ , the root  $r$  and mutation rate  $\theta$ , and also includes the temporal ordering  $\mathcal{H}$  as a latent variable. In order to draw samples from the posterior distribution of the parameters of interest  $p(r, T | \mathcal{S})$ , we construct a Markov chain Monte Carlo sampler. The chain





**FIG. 7. (Top)** In this tree the MRCA of the sample (**top**) is observed twice in the sample. Note that one of the intermediate haplotypes is not observed in the sample (and hence has zero sample size). **(Bottom)** A possible scenario for how the present sample came about. Nodes without a number represent haplotypes that have not arisen yet. At first one sequence is present, the ancestral sequence, which split into two (remember that the first event is always a split). Then one of those two identical sequences split again to give us a total of three. One of those three then mutates to give rise to the intermediate haplotype, which in turn splits 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 subsequently splits to give another copy of itself.

is initialized by drawing a mutation rate  $\theta^{(0)}$ , generating a tree  $T^{(0)}$ , and picking root  $r^{(0)}$  uniformly from  $T^{(0)}$ .

1. Propose a new root by using the prior distribution as a proposal kernel over all available sequences  $q(r \rightarrow r') = p(r')$ , and sample a latent temporal ordering  $\mathcal{H}' = \{\mathcal{H}'_1, \dots, \mathcal{H}'_j\}$  according to Algorithm 1.1 with probability  $q(\mathcal{H}'|T, r')$ . Accept or reject  $(r', \mathcal{H}')$  according to the corresponding Metropolis-Hastings ratio  $\min(1, A_r)$ , where

$$A_r = \frac{P(\mathcal{H}'|\mathcal{S}, H, r', \theta)}{P(\mathcal{H}|\mathcal{S}, H, r, \theta)} \times \frac{q(\mathcal{H}|T, H, r, \theta)}{q(\mathcal{H}'|T, H, r', \theta)}$$

2. Propose a new tree topology  $T'$  at random (implying a number of events  $H'$ ), and sample  $\mathcal{H}' = \{\mathcal{H}'_1, \dots, \mathcal{H}'_j\}$  according to Algorithm 1.1 with probability  $q(\mathcal{H}'|T', r)$ . Accept or reject the new tree topology and latent ordering according to the corresponding Metropolis-Hastings ratio  $\min(1, A_T)$ , where

$$A_T = \frac{P(\mathcal{H}'|\mathcal{S}, H', r, \theta)}{P(\mathcal{H}|\mathcal{S}, H, r, \theta)} \times \frac{q(\mathcal{H}|T, H, r, \theta)}{q(\mathcal{H}'|T', H', r, \theta)}.$$

3. Propose new mutation rate from the prior  $\theta \sim IG(a_\theta, b_\theta)$ , and accept according to the corresponding Metropolis-Hastings ratio  $\min(1, A_\theta)$ , where

$$A_\theta = \frac{P(\mathcal{H}|\mathcal{S}, H, r, \theta')}{P(\mathcal{H}|\mathcal{S}, H, r, \theta)}.$$

### C. Generalized time-reversible mutation model

Here we present the most general time-reversible mutation model under assumptions of independent sites, stationarity of nucleotide frequencies, time-reversibility and no selection, namely the Generalised Time-Homogeneous Time-

Reversible model (GTR) (Tavaré, 1986). We consider  $l$  (the length of the sequences) parallel independent mutation processes and represent the state of each nucleotide site  $j$  of sequence  $i$  as  $X_j^i$ . Mutations occur as a Markov Process with generator Q-matrix

$$Q = \phi_j \begin{pmatrix} \cdot & v_1\pi_G & v_2\pi_C & v_3\pi_T \\ v_1\pi_A & \cdot & v_4\pi_C & v_5\pi_T \\ v_2\pi_A & v_4\pi_G & \cdot & v_6\pi_T \\ v_3\pi_A & v_5\pi_G & v_6\pi_C & \cdot \end{pmatrix}$$

where the  $\pi_s$  ( $s = A, G, C, T$ ) represent the equilibrium probabilities of the nucleotides, and the mutation coefficients  $v_1, \dots, v_6$  the relative mutation probabilities. The diagonal elements are such that each row sums to zero. The extra parameter  $\phi_j$  denotes the site-specific mutation rate for each site  $j$ . A Markov process at time  $t$  with generator matrix  $Q$  and initial distribution equal to the distribution  $\delta_i$  (here  $\delta$  is the Kronecker  $\delta$  and represents a known initial state  $i$  at time 0) can be viewed as a Markov chain with transition matrix

$$P_{ik}^{(t)} = \{\exp(Qt)\}_{ik}.$$

This implies that mutations happen as a Poisson Process with rate  $\phi_j q_i$  when at each state  $i$  in site  $j$ , where  $q_i$  is the sum along row  $i$  of the rates of jumping to other possible states.

## DISCLOSURE STATEMENT

No competing financial interests exist.

## REFERENCES

- Bahlo, M., and Griffiths, R. 2000. Inference from gene trees in a subdivided population. *Theor. Popul. Biol.* 57, 79–95.
- Beaumont, M. 2003. Estimation of population growth or decline in genetically monitored populations. *Genetics* 164, 1139–1160.
- Bloomquist, E.W., Lemey, P., and Suchard, M.A. 2010. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol. Evol.* 25, 626–632.
- Crandall, K., and Templeton, A. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny reconstruction. *Genetics* 134, 959–969.
- De Iorio, M., and Griffiths, R. 2004. Importance sampling on coalescent histories. *I. Adv. Appl. Probabil.* 36, 417–433.
- Emerson, B., and Hewitt, G. 2005. Phylogeography. *Curr. Biol.* 15 R367–R371.
- Ethier, S. N., and Griffiths, R. C. 1987. The infinitely-many-sites model as a measure-valued diffusion. *Ann. Probabil.* 15, 515–545.
- Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410.
- Griffiths, R., and Tavaré, S. 1994. Ancestral inference in population genetics. *Stat. Sci.* 9, 307–319.
- Hewitt, G. 2000. The genetic legacy of the Quaternary ice ages. *Nature* 405, 907–913.
- Hudson, R., and Kaplan, N. 1988. The coalescent process in models with selection and recombination. *Genetics* 120, 831–840.
- Kingman, J.F.C. 1982. The coalescent. *Stochastic Process. Appl.* 13, 235–24
- Lemey, P., Rambaut, A., Drummond, A.J., et al. 2009. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* 5, e1000520.
- Lemey, P., Rambaut, A., Welch, J.J., et al. 2010. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* 27, 1877.
- Manolopoulou, I., 2009. A Bayesian approach to nested clade analysis. [Ph.D. dissertation]. University of Cambridge, Cambridge, UK.
- Manolopoulou, I., Legarreta, L., Emerson, B. C., et al. 2011. A Bayesian approach to phylogeographic clustering. *J.R. Soc. Interface Focus* 1, 909–921.
- Neuhauser, C., and Krone, S. 1997. The genealogy of samples in models with selection. *Genetics* 145, 519–534.
- Posada, D., and Crandall, K. 2001. Intraspecific gene genealogies: trees grafting into networks. *Trends Ecol. Evol.* 16, 37–45.
- Slatkin, M. 2001. Simulating genealogies of selected alleles in a population of variable size. *Genet. Res.* 78, 49–57.
- Stephens, M., and Donnelly, P. 2000. Inference in molecular population genetics. *J. R. Stat. Soc. Ser. B Methodol.* 62, 605–655.

- Swofford, D.L., and Berlocher, S.H. 1987. Inferring evolutionary trees from gene frequency data under the principle of maximum parsimony. *Syst. Biol.* 36, 293.
- Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* 17, 57–86.
- Tavaré, S. 2003. *Nature Encyclopedia of the Human Genome*. Nature Publishing Group, London.
- Wakeley, J. 2008. *Coalescent Theory: An Introduction*. Roberts and Company Publishers, Greenwood Village, CO.

Address correspondence to:  
*Dr. Ioanna Manolopoulou*  
*Department of Statistics*  
*Duke University*  
*Box 90251*  
*Durham, NC 27708*

*E-mail: ioanna.manolopoulou@duke.edu*