Divergence estimation in the presence of ILS and migration

Graham Jones

2017-05-30, June 18, 2017

This is a work in progress...

We describe a BEAST2 package DENIM for species tree estimation using the multispecies coalescent, in the presence of some migration. It is based on an approximation which breaks down if there is a lot of migration. Leaché et al. (2014) showed migration causes problems for species tree inference using the multispecies coalescent when migration is present but ignored.

1 Introduction

'Migration' is used to refer to gene flow between species (usually introgression but not restricted to that). We use the term 'species' rather than 'population' because the method is aimed at situations where gene flow is small. A migration event occurs when an allele comes from a parent from another species. An 'embedding' of a gene tree specifies which species tree branch each coalescence belongs to, together with migration events, which specify the times along gene tree branches at which an allele moved between species tree branches, and which species tree branch is the destination. We always describe events going back in time from the present, so alleles have parental species to which they 'go', and the 'destination' branch in the species tree contains part of a gene tree branch at a more ancient time than the 'source' branch. This is because coalesences are easier to model this way, and is the same convention as the program IMa2 (Hey and Nielsen, 2004, 2007; Hey, 2010).

TODO. Mention phylogenetic networks, Nakhleh, Zhang et al (inc Stadler) 2017, Solís-Lemus and Ané (2016). Here we use species trees with some gene flow, but not enough to make a tree model unsuitable.

There is no upper limit on the number of migration events, and even if this is limited, and the gene tree and species tree are fixed, there can be a huge number of ways in which each gene tree can be embedded into a species tree. It is thus difficult to make inferences if the situation modelled in full. IMa2 requires that the true population phylogeny (equivalent to species tree here) is known.

We use a model for migration which is similar to that used by Hey (2010) in IMa2. There is a migration rate parameter for each ordered pair of contemporaneous species tree branches. There are $2(n-1)^2$ of them for a species tree with n tips (Hey, 2010). There are three main differences between DENIM and IMa2. We estimate the species tree instead of assuming it; we integrate out the migration rate parameters; and we use an approximation to simplify sampling from the posterior. We also integrate out the population size parameters in a similar fashion to Jones (2016). Our focus is on estimating the species tree despite the presence of small amounts of migration. Since migration rates are integrated out, they cannot be estimated, but DENIM does estimate the number of migrations within each locus.

Even with migration rate and population size parameters integrated out, there are still an unbounded number of parameters for the gene trees. It appears very difficult to design and implement MCMC operators capable of sampling efficiently from this distribution while estimating the species tree. Here we use an approximation to the posterior by ignoring most of the 'unlikely' embeddings. If the migration rate is high, some of the ignored embeddings will be quite likely and the approximation will break down.

Difficulties: Sousa et al. (2011) On the non-identifiability of migration time estimates in isolation with migration models, Hey et al. (2015) On the occurrence of false positives in tests of migration under an isolation-with-migration model

Maybe cite Palczewski and Beerli (2013), A Continuous Method for Gene Flow, (approx for high rates).

2 The prior density for a gene tree

2.1 Background

Following the introduction of the Kingman coalescent (Kingman, 1982), models for coalescence and migration were developed in the 1980s by population geneticists (Hudson et al., 1990). More recent developments include Beerli and Felsenstein (2001), Ewing and Allen (2006), Tian and Kubatko (2016), Dalquen et al. (2016) as well as the work of Hey and Nielsen. The methods of Tian and Kubatko (2016) and Dalquen et al. (2016) can estimate the species tree, but are currently restricted to at most 3 species and 3 sequences per locus.

The underlying evolutionary model we use here is the same as that of Hey (2010), except that the species tree S is not assumed known but instead follows a birth-death model. When the species tree S is estimated, it is important that $\int \Pr(G|S)dG = 1$ for any S, where G is a gene tree. I have not found a clear statement to this effect in the literature, so some explanation seems warranted. Between the node heights of the species tree, we have an n-island model for coalescence and migration (Beerli and Felsenstein, 2001), where n is the current number of species tree branches. This is a continuous time Markov chain. It could be time-inhomogenous, to allow for population sizes or migration rates to vary continuously with time, although our application here only uses the time-homogenous case. In order to define the state space of this Markov chain, we need a few preliminaries.

Firstly, each branch in G is labelled by the tip labels that descend from the branch. When a coalescence occurs, it should be understood as the merging of two particular labelled gene tree branches. Likewise, when a migration occurs, a particular gene tree branch migrates to a particular species tree branch. Let L be the set of tip labels of G, and let $\mathcal{P}(L)$ be the set of all partitions of L. Each partition $P \in \mathcal{P}(L)$ is a set $\{L_1, \ldots, L_m\}$ for some m with $1 \leq m \leq |L|$, where each L_i is a nonempty subset of L, the union of them all is L, and they are pairwise disjoint. The subsets L_i are called the 'blocks' of the partition. At any time, the set of gene tree branches can be regarded as a member of $\mathcal{P}(L)$, and each branch as a block. We will call the periods between node heights of S, during which the number of branches is constant, an 'epoch'. The branches of S could be labelled in a similar manner to G, but for convenience, we assume they have been labelled with the numbers $\{1, \ldots, n\}$ during the epoch when there are n branches, and that branches n and $n-1$ merge to form a branch $n-1$ in the next epoch.

The state space of the Markov chain during the epoch with n branches consists of all possible assignments of all members of $\mathcal{P}_n(L)$ to the branches of S. Each state is a pair (P, f) where $P \in \mathcal{P}_n(L)$ and f is any map from P to $\{1, \ldots, n\}$, assigning gene tree branches to species tree branches. We use the set theory notation X^Y to denote the set of all maps from set Y to set X. So we can write the state space $\mathcal{A}_n$ as

$$\mathcal{A}_n = \{(P, f) \ : \ P \in \mathcal{P}(L) \ \wedge \ f \in \{1, \ldots, n\}^P\}.$$

It has size

$$|\mathcal{A}_n| = \sum_{P \in \mathcal{P}(L)} n^{|P|}.$$

There is an instantaneous rate matrix Q of size $|\mathcal{A}_n| \times |\mathcal{A}_n|$. The off-diagonal rows of Q are non-negative, the rows of Q sum to zero, and the diagonal entries are less than or equal to zero. In fact all the diagonal entries are strictly negative, except that $Q_{z,z} = 0$ where z is the final state in the root of the species tree, when $n = 1$, and there single gene tree branch. Note that although Q is enormous for large $|L|$ and n, it is extremely sparse, since the number of states which can be reached from a given state by a single migration or coalesence is much smaller than $|\mathcal{A}_n|$. Basic properties of Markov chains (in particular the fact that rows of Q sum to zero) ensure that given a starting distribution over states such that

$$\sum_{(P,f) \in \mathcal{A}_n} \Pr(P, f) = 1,$$

this remains true throughout the process, and in particular just before a merging of species tree branches. At such a merge, the partitions P are unchanged, but the state space changes.

Once we are in the root branch of the species tree, the process reduces to the Kingman coalescent, which is a (normalised) density. Consider the case just above the root, where $n = 2$. We have

$$\sum_{(P,f) \in \mathcal{A}_2} \Pr(P, f) = \sum_{P \in \mathcal{P}(L)} \sum_{f \in \{1,2\}^P} \Pr(P, f).$$

Each P consist of blocks $L_1, \ldots, L_m$, and as f runs over the maps from P to $\{1, 2\}$, it runs over exactly those assignments of these blocks to $\{1, 2\}$ which result in all of them ending up in the root just after the merge. Thus

$$\sum_{f \in \{1,2\}^P} \Pr(P, f) = \sum_{f \in \{1\}^P} \Pr(P, f)$$

where the left hand side applies just before the merge and the right hand side applies just after the merge. It follows that

$$\sum_{(P,f) \in \mathcal{A}_2} \Pr(P, f) = \sum_{(P,f) \in \mathcal{A}_1} \Pr(P, f)$$

where again the left hand side applies just before and the right hand side applies just after the merge. We can apply a similar argument to merges when $n > 2$ to establish that $\int \Pr(G|S) dG = 1$. We will refer to this this evolutionary model as the 'tree-island model'.

2.2 Integrating out population and migration parameters

Suppose the species tree has s tips. There are $2s - 1$ species tree branches, including the root branch. Suppose the migration rate from branch b to branch d is m_{bd}. These migration rates follow the same conventions as Hey and Nielsen (backwards from present, scaled by population size). Population size parameters θ_j and ploidy values p_j are as Jones (2016).

The time (going back from zero at present) is divided into a number of intervals τ_i ($i \in \mathcal{I}$) by the times of the events and species tree node heights. The set of species tree branches which exist during the ith interval is denoted by $\mathcal{B}_i$, and we set $s_i = |\mathcal{B}_i|$. The number of lineages in gene tree j which belong to species tree branch b during the ith interval is n_{jbi}. The set of intervals which end in a coalesence is $\mathcal{I}_{coal}$, and the set which end in a migration is

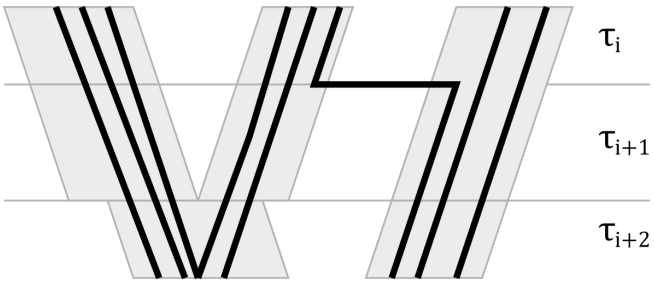

Figure 1: Three time steps. The first ends in a migration, the second with a species tree node, and the third with a coalescence.

$\mathcal{I}_{mig}$. See Figure 1, where $i \in \mathcal{I}_{mig}$, $i+1$ is in neither, and $i+2 \in \mathcal{I}_{coal}$. The rate at which the next event occurs is $(\kappa_i + \mu_i)$ where

$$\kappa_i = \sum_j \sum_{b \in \mathcal{B}_i} \left(\binom{n_{jbi}}{2} p_j^{-1} \theta_b^{-1} \right)$$

is the total rate for coalescent events and

$$\mu_i = \sum_j \sum_{b \in \mathcal{B}_i} \left(n_{jbi} \sum_{d \in \mathcal{B}_i \setminus b} m_{bd} \right) \tag{1}$$

is the total rate for migration events. Here we are summing the nonzero off-diagonal elements of a row of Q in order to find $Q_{x,x} = -(\kappa_i + \mu_i)$ for the current state x. We then need $Q_{x,y}$ where y is the next state. If it is a coalescence, $Q_{x,y} = p_{j_i}^{-1} \theta_{b_i}^{-1}$, where j_i is the gene tree containing the coalescence, and b_i is the species tree branch in which it occurs. If it is a migration, $Q_{x,y} = m_{b_i d_i}$, where b_i and d_i are the source and destination branches.

Denoting the set of migration rates by M and the set of population size parameters by Θ, we have

$$f(G; S, \Theta, M) = \prod_{i \in \mathcal{I}_{coal}} p_{j_i}^{-1} \theta_{b_i}^{-1} \prod_{i \in \mathcal{I}_{mig}} m_{b_i d_i} \prod_i \exp(-(\kappa_i + \mu_i)\tau_i)$$

This can be factored into a coalescence part and a migration part. Then, our aim is to rearrange the terms in the coalescence part so that it is a product over species tree branches, and the rearrange the terms in the migration part so that it is a product over pairs of species tree branches. The result will be a product of terms, in which each term contains one populaton size parameter θ_b or one migration parameter m_{bd}. This enables us to integrate out these parameters if suitable priors are assumed. We put

$$f(G; S, \Theta, M) = f_{coal}(G; S, \Theta) \;\; f_{mig}(G; S, M)$$

where

$$f_{coal}(G; S, \Theta) = \prod_{i \in \mathcal{I}_{coal}} p_{j_i}^{-1} \theta_{b_i}^{-1} \;\; \exp\left(-\sum_i \tau_i \kappa_i \right)$$

and

$$f_{mig}(G; S, M) = \prod_{i \in \mathcal{I}_{mig}} m_{b_i d_i} \;\; \exp\left(-\sum_i \tau_i \mu_i \right) \tag{2}$$

First we deal with f_{coal}. We have

$$\prod_{i \in \mathcal{I}_{coal}} p_{j_i}^{-1} \theta_{b_i}^{-1} = \prod_j \prod_b (p_j \theta_b)^{-k_{jb}}$$

where k_{jb} is the number of coalescences in gene tree j in branch b. Next

$$\sum_i \tau_i \kappa_i = \sum_i \tau_i \sum_j \sum_{b \in \mathcal{B}_i} \binom{n_{jbi}}{2} p_{j_i}^{-1} \theta_b^{-1} = \sum_b \sum_j \sum_{i: b \in \mathcal{B}_i} \binom{n_{jbi}}{2} p_j^{-1} \tau_i \theta_b^{-1}$$

so

$$f_{coal}(G; S, \Theta, M) = \prod_b r_b \theta_b^{-q_b} \exp\big(-\gamma_b \theta_b^{-1}\big), \quad \text{where}$$

$$q_b = \sum_j k_{jb}, \qquad r_b = \prod_j p_j^{-k_{jb}}, \qquad \text{and} \quad \gamma_b = \sum_j \sum_{i: b \in \mathcal{B}_i} \binom{n_{jbi}}{2} p_j^{-1} \tau_i. \tag{3}$$

As written, there are time intervals in branch b for events during which no change occurs in branch b. For the computation of γ_b, we only need to take into account coalescences within branch b and migrations in and out of branch b, since between these events, n_{jbi} is constant. Equation (3) is now of the same form as equation (2) of Jones (2016). The only difference is that γ_b accounts for migrations in and out of branch b. This means the population size parameters can be integrated out as in Jones (2016).

Now we turn to the migration part f_{mig}. Let $\mathcal{O}$ be the set of contemporaneous pairs of branches in S. We have

$$\sum_i \tau_i \mu_i = \sum_i \tau_i \sum_j \sum_{b \in \mathcal{B}_i} n_{jbi} \sum_{d \in \mathcal{B}_i \setminus b} m_{bd} = \sum_{(b,d) \in \mathcal{O}} \sum_{i: b,d \in \mathcal{B}_i} \tau_i \sum_j n_{jbi} m_{bd}$$

Thus

$$f_{mig}(G; S, \Theta, M) = \prod_{(b,d) \in \mathcal{O}} m_{bd}^{n_{bd}} \exp\left(-\zeta_{bd} m_{bd}\right)$$

where n_{bd} is the total number of migrations from b to d and

$$\zeta_{bd} = \sum_{i: b,d \in \mathcal{B}_i} \tau_i \sum_j n_{jbi}. \tag{4}$$

The term ζ_{bd} can be interpreted as the total intensity of migrations from b to d during the time in which both branches b and d exist. If we assume that $m_{bd} \sim \mathcal{G}(\alpha_{bd}, \beta_{bd})$ for all b, d where $\mathcal{G}$ is the gamma distribution, then we get a contribution to the posterior which is

$$\begin{aligned} & \prod_{(b,d) \in \mathcal{O}} \int_0^\infty \frac{\beta_{bd}^{\alpha_{bd}}}{\Gamma(\alpha_{bd})} m_{bd}^{\alpha_{bd}} \exp(-\beta m_{bd})\ m_{bd}^{n_{bd}} \exp\left(-\zeta_{bd} m_{bd}\right) \mathrm{d}m_{bd} \\ = & \prod_{(b,d) \in \mathcal{O}} \frac{\Gamma(n_{bd} + \alpha_{bd})}{\Gamma(\alpha_{bd})} \frac{\beta_{bd}^{\alpha}}{(\beta_{bd} + \zeta_{bd})^{n_{bd} + \alpha_{bd}}} \end{aligned} \tag{5}$$

Equations (4) and (5) provide the information needed to implement the calculation for the migration part of the posterior. We have allowed each ordered pair of contemporaneous branches (b, d) to have a different prior. For example, we can represent the prior expectation that migration rates are lower between more distantly related branches.

The calculation in equation (4) is slow when the number of tips in the species tree is large. A much simpler model is is to assume that m_{bd} is the same value m for all b, d. In this case, equation (1) reduces to

$$\mu_i = m \sum_j \sum_{b \in \mathcal{B}_i} n_{jbi}(s_i - 1)$$

The double sum is equal to the total number of gene tree lineages N_i during time interval i. Then we have

$$f_{mig}(G; S, M) = m^N \exp\Big(-\sum_i \tau_i \mu_i \Big) = m^N \exp\Big(-m \sum_i \tau_i N_i (s_i - 1) \Big).$$

where N is the total number of migrations. The parameter m could be integrated out or sampled during the MCMC.

TODO specify exact models used somewhere.... Here or start of results?

3 How the gene tree is embedded

This section describes the embedding parameters, and how they are used to embed the gene trees. We restrict the embeddings by ignoring ones which are unlikely when the migration rates are small enough. The hope is that we will still explore a region of parameter space which includes most of the probability content.

Embeddings are restricted by applying the following rules:

1. there is at most one migration in a single gene tree branch
2. at most one of the child branches of a gene tree node contains a migration
3. there are no more migrations than needed (in a sense described below)

We call a pair of child branches of a gene tree node a **sister-pair**. The embedding parameters E_j consist of two values $\xi_{ji}, \eta_{ji} \in [0, 1]$ for each internal node i of the jth gene tree. See Figure 2. The parameter ξ_{ji} determines where along a sister-pair a migration may occur, if a migration is needed in th embedding. Thus it determines which child branch of node i is capable of migrating, as well as the time of the migration if there is one. All the nodes in the species tree have their children labelled as 'left' and 'right', so that $[0, 1]$ can be mapped unambiguously onto the sister-pair. The other parameter η_{ji} specifies which of the node's child branches to use when choosing a destination species branch for an introgression. If the migration is between sister branches of the species tree, there is only one choice for the destination. It may happen that the sister branch is too ancient, in which case several destination species branches are possible. The possible destination branches are found, and η_{ji} is used to choose between them by dividing the interval [0,1] equally into the appropriate number of parts.

The parameters ξ_{ji} and η_{ji} are changed by operators during the MCMC algorithm, regardless of whether or not they are being used to embed a gene tree. This is a simpler alternative to implementing rjMCMC operators which account for changes in dimension. The prior $\Pr(E_j)$ for E_j is are independent uniform distributions on $[0, 1]$ for each ξ_{ji} and η_{ji}.

The first rule above is straightforward. The definition of ξ_{ji} enforces the second rule. The third rule is applied recursively from the tips. Suppose x is the ith node of the jth gene tree, and suppose both child nodes of x have been assigned to branches in the species tree. If it is possible to assign x to a species tree branch without an migration in either child branch of x, then this is done. Otherwise x is assigned using the species tree branch to which its non-migrating child has been assigned: it will be the same branch, or an ancestor of that branch,

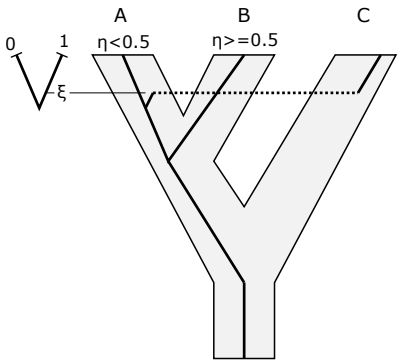

Figure 2: This shows ξ_{ji} and η_{ji} for a one gene tree and one node (with subscripts dropped). There is a migration from C into A. The parameter ξ determines how far along the sister-pair the migration occurs, and η determines whether the destination branch is A or B.

depending on the height of x. The height of the migration is fixed by ξ_{ji}. The migrating child branch of x starts in the species tree branch that the migrating child has been assigned. It stays in this branch, or an ancestor of it, until the migration height. It will then migrate to the same species tree branch as x, or a descendant of it. If there is more than one descendant of the species tree branch of x at this height, values from η_j are used to choose one.

3.1 Properties of the embedding scheme

Different embeddings of the same gene tree in the same species tree are obtained by changing ξ_{ji} and η_{ji} during the MCMC sampling. Figure 3 shows some examples. Case (a) is simple. No migrations are needed to embed the gene tree, so embeddings with one or more migrations are ignored. Case (b) requires one migration, and an embedding with two migrations in the same branch is ignored. Case (c) requires two migrations. The embedding on the left is ignored since it has two sister branches with migrations. The embedding on the right is one of four embeddings that is considered.

Proposition *Given any set of particular values for ξ and η, and the rules above, any gene tree can be embedded in any species tree. For any G_j and S, the set of embeddings as ξ and η vary include at least one with a minimal number of migrations.*

Proof: The first claim is straightforward, using recursion starting at the tips, and following the description above (for applying the third rule).

For second claim, suppose it is false and consider the set M of minimal embeddings (those with a minimal number of migrating branches). Call a node both of whose child branches migrate a 'double node'. Thus every member of M has at least one double node. Now restrict attention to the subset $\bar{M}$ of M of embeddings which have as few as possible double nodes. Finally, choose an embedding B from $\bar{M}$ so that a double node x is as near to the root as possible.

If x is the root, it can be moved into the same branch as one of its children, or an ancestor of that branch, and one migration can be removed, contradicting the definition of M. If x is not the root, it can be again moved into the same branch as one of its children, but now the branch between x and its parent may need to become migrating. If the sister branch to x is already migrating, we have an embedding with the same number of migrations, but a

double node closer to the root than x, contradicting the definiton of B. If the sister branch to x is not migrating, we have an embedding with fewer double nodes than B, contradicting the definition of $\bar{M}$. End of proof.

The method does not consider every embedding which has a minimal number of migrations (eg Figure 3c). Some embeddings which are considered are not minimal. Eg gene tree ((a1,b1),b2) with b2 and b2 in same species, and a species tree with large root height. If (a1,b1) is assigned to the branch a1, two migrations will be used, but it is possible to use only one.

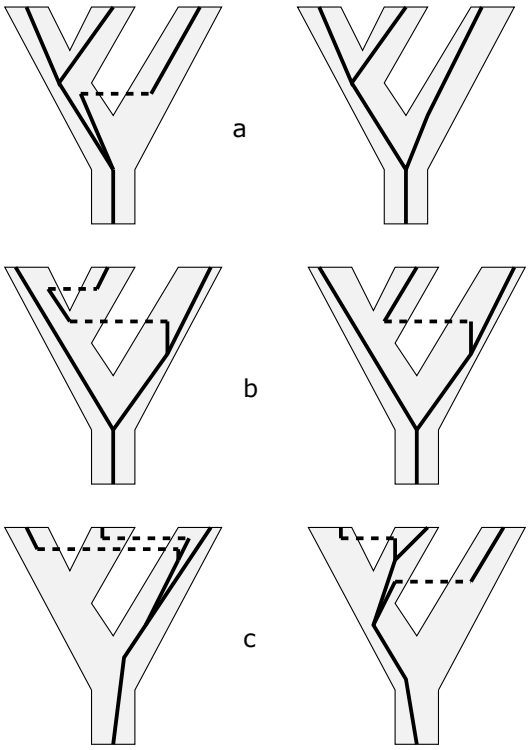

Figure 3: Some examples of embeddings for three gene trees in a, b, c. Branches which introgress are in red. On the left are embeddings that are ignored. On the right is an embedding which is considered.

4 Implementation notes

DENIM uses the standard tree operators implemented in BEAST2 for the species tree and the gene trees. For a standard multispecies coalescent analysis, operators which change the species tree and the gene trees in a coordinated way are beneficial (Jones, 2016; Ogilvie et al., 2017). These operators rely on, and preserve, compatibility between the species tree and gene trees under the multispecies coalescent. In the presence of migration, any gene tree is compatible with any species tree, and these co-ordinated operators cannot be used as they are.

A couple of simple MCMC operators were implemented for the embedding parameters. As noted above, they are changed by operators regardless of whether or not they are being used to embed a gene tree. In general, applying

MCMC operators to unused parameters could be very inefficient, and rjMCMC would be preferable, but here the operators for ξ and η are very fast.

DENIM is implemented in the BEAST2 framework, and so benefits from the flexible site models, substitution models, and others available in BEAST2. An analysis can be set up using the graphical interface Beauti.

TODO: Explain `Relatedness Factor` and `Migration Decay Scale`.

5 Some results with prior only

TODO: Not sure what is useful here...

6 Results on simulated data

TODO: 10 species.

TODO: More replicates.

TODO: More experiments with flexible model and different prior means for migration rate.

TODO: Some results will go into SI...

This uses some data from Leaché et al. (2014). For now, there are 4 species (A,B,C,D), and twenty replicates. The settings for site and clock models were similar to those used by Leaché et al. (2014). Site models were linked. A GTR model of substitution was used, with base frequencies equal. The clock models were strict but unlinked. The first locus had clock rate 1, and the others were estimated. The Yule (pure birth) model was used for the species tree. The priors were set as follows. Substitution rates relative to rateCT: Gamma(0.05,20). Relative clock rates: lognormal(0.1). Growth rate for the species tree: lognormal(5,2). PopPriorScale: lognormal(-5,2).

The results here use the simple model for migration, with an exponential prior with mean 0.001.

6.1 Coverage

Coverage means the number of times that the correct species tree topology appears in the 95% credible set. Results are shown in the table.

Rate	0.001	0.01	0.1	1.0
IM	20	20	20	20
Ancestral-IM	20	20	19	20
Paraphyly	20	19	4	4
No migration	20			
Sister-1-allele	20			
Sister-1-mig	20			
Nonsister-1-allele	19			
Nonsister-1-mig	17			

For the two migration patterns with migration between sister species, the coverage was almost complete, in line with results from Leaché et al. (2014). DENIM succeeded in the paraphyly migration pattern (where there is

migration between branches B and C) when the migration rate is 0.001 or 0.01, and breaks down for 0.1 and 1.0. DENIM also succeeded in the cases where a single allele or migrant moves between species at time zero.

6.2 Branch scores

The tree distance used is the branch score of Kuhner and Felsenstein (1994), adapted for rooted trees. It accounts for differences in topology and branch lengths. The entire posterior is evaluated by finding the mean distance between the MCMC samples of the species tree and the true tree. Figure 4 shows the results. The mean distances are low for migration rates of 0.001 and 0.01, in the cases where a single allele or migrant moves between species at time zero. They become large for the higher migration rates, including the cases of migration between sister species where the toplogy is correct.

6.3 Migration estimates

Figure 5 shows how DENIM infers the existence of migrations. They show that DENIM very rarely infers migration when there is none. It can detect migration quite reliably when it is between non-sister species, and is at a small rate. For migration between sister species the performance is mixed.

7 Results on empirical data

We re-analysed the pocket gopher data of Belfiore et al. (2008). We used the HKY substitution model, linked site models, estimated relative clock rate for all loci except the first, and a strict clock. The results here use the simple model for migration, with an exponential prior with mean 0.001. (TODO also tried 0.0002, with similar results, but lower PP 0.82 for the ingroup clade. Presumably making it small enough will cause the outgroup to be misplaced.)

This data was also analysed in Heled and Drummond (2010), the paper which introduced *BEAST. In the *BEAST analysis, the outgroup species *Orthogeomys heterodus* was misplaced (their Figure 8a) , and the authors comment that "The tendency to place the outgroup incorrectly appears to be caused by just one gene" namely TBO29. The tree from the DENIM analysis is shown in Figure 6. The outgroup is correctly placed, and it is very similar to the *BEAST result with ingroup monophyly enforced (their Figure 8b). The DENIM tree is somewhat shorter, perhaps due to a different site model or population model. A migration was inferred between *Orthogeomys heterodus* and the (*T. bottae, Thomomys townsendi, Thomomys umbrinus*) clade, the same clades that *BEAST grouped together. This migration was present in about 95% of the MCMC samples. The other migrations that appear in the posterior samples have much lower posterior probabilities. The next migrations that DENIM analysis suggests (at about 24%) are very recent ones, both ways, of TBO47 between *T. bottae* and *T. umbrinus*. This pair is followed (at about 18%) by a very recent one of TBO64 *T. talpoides* to *T. idahoensis* (going back in time).

It is interesting that DENIM identifies TBO64, but not TBO29, as a locus with migration. The posterior mean count of migrations for TBO64 was 1.20, for TBO47 it was 0.47, for TBO26 it was 0.11, and the rest, including TBO29, were well under 0.1. In Belfiore et al. (2008), the individual gene trees were estimated separately, and it appears from their Figure 2 that in TBO64, the relative distance between *Orthogeomys heterodus* and other taxa is considerably smaller than is the case for any other locus.

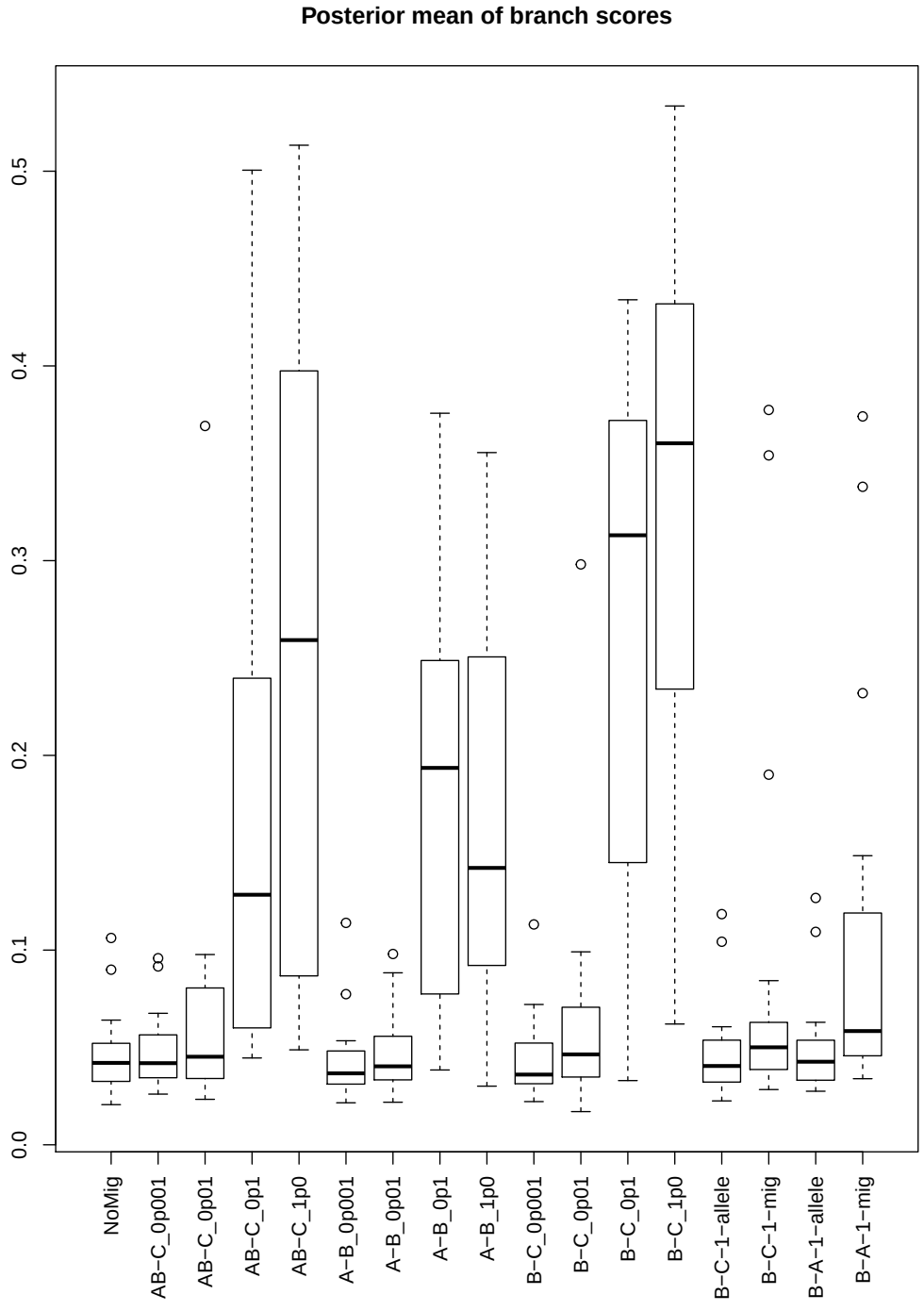

Figure 4: Branch scores.)

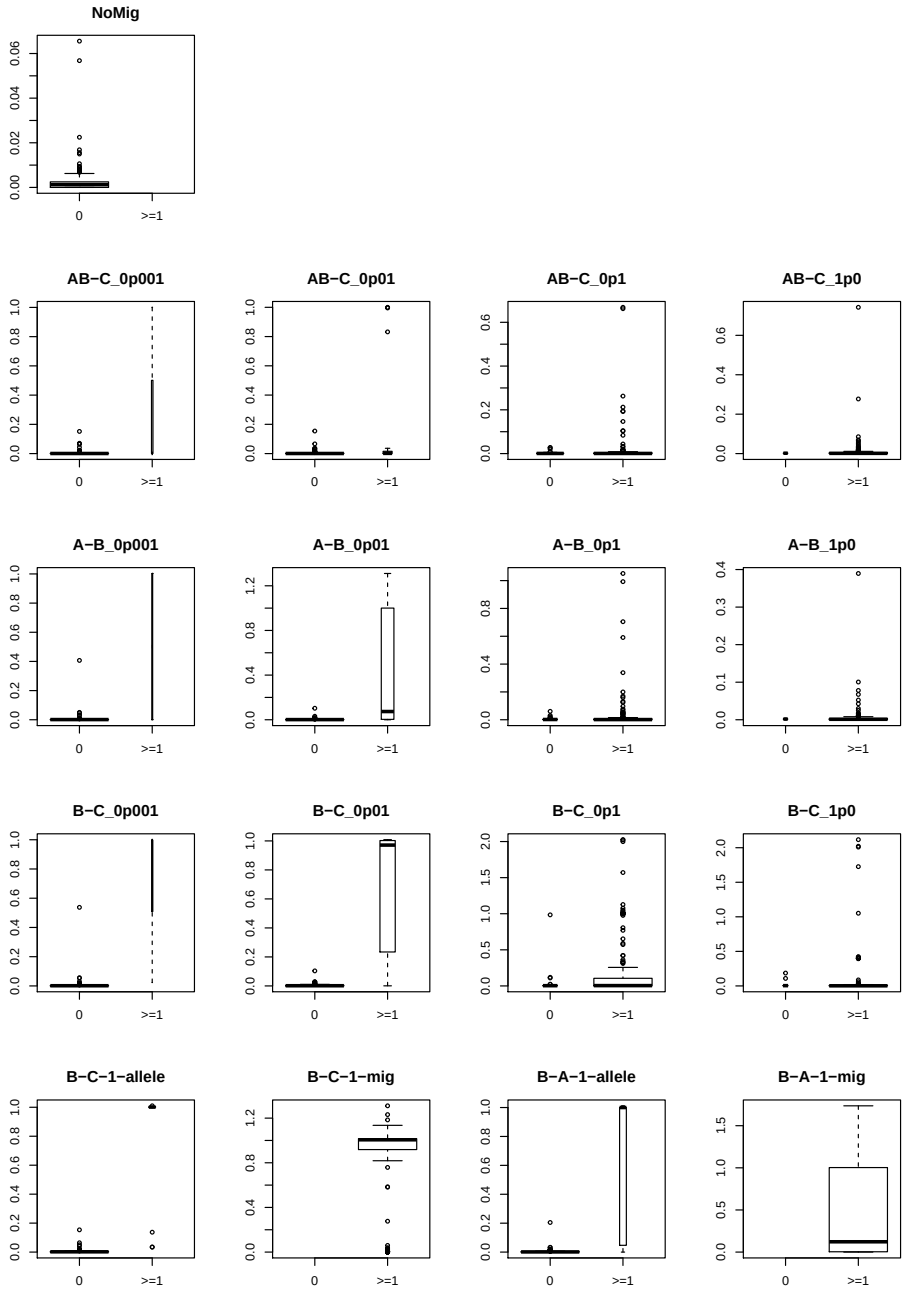

Figure 5: Posterior mean counts of migration, divided into two cases for each migration pattern and migration rate. Each boxplot is based on a total of 20 replicates and 10 loci. The label `0` means there is no migration at a locus in the simulated data, and the label `>=1` means there is some migration at a locus in the simulated data. The widths of the boxes indicate the proportions belonging to the two cases.

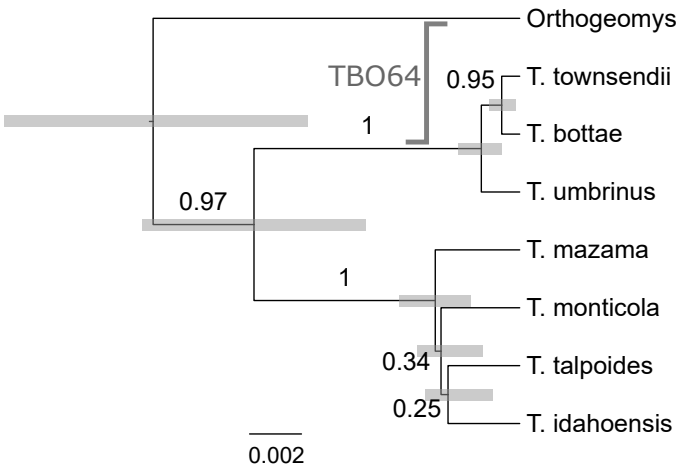

Figure 6: Gopher tree. Posterior clade probabilities are shown next to branches. The node bars are 95% HPDs for the node heights. The migration of of locus TBO64 is also indicated.

8 Discussion

The tree-island model is independent of the approximation used in DENIM. The partial sampling of the posterior is a trade-off between accuracy on the one hand and computational effort and simplicity of implementation on the other. TODO.

Species delimitation. TODO.

Small amounts of paraphyletic migration can be very disruptive to species tree estimation, and DENIM is able to deal with these cases effectively. In these cases, the migrating loci can often be identified. Results for migration patterns only involving migration between sisters are similar to *BEAST. In these cases, the migrating loci are hard to identify.

The program can identifies loci which are 'badly behaved', rather than those which migrate. That is, it identifies loci with migrations which result in an incompatibility with the species tree. Some migrations do not cause incompatibility, because (going back in time) they do not coalesce with another lineage until the species tree branches have merged. In other cases, a lineage may migrate, then migrate back again before coalescing, or two lineages may both migrate to the same species branch, coalesce there, and then not coalecse with other lineages until the species tree branches have merged. There are other, less likely, situations where migration does not result in incompatibilities.

TODO cite Sousa about unidentifiable migration time estimates.

Birth-death process instead of pure birth. Extinction. If there are extinct species, migration can result in unusually deep coalescences. Could infer. TODO

References

Peter Beerli and Joseph Felsenstein. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of the National Academy of Sciences*, 98 (8):4563–4568, 2001.

Natalia M. Belfiore, Liang Liu, and Craig Moritz. Multilocus phylogenetics of a rapid radiation in the genus thomomys (rodentia: Geomyidae). *Systematic Biology*, 57(2):294, 2008. doi: 10.1080/10635150802044011. URL + `http://dx.doi.org/10.1080/10635150802044011`.

Daniel A. Dalquen, Tianqi Zhu, and Ziheng Yang. Maximum likelihood implementation of an isolation-with-migration model for three species. *Systematic Biology*, 00:00–00, 2016.

Greg Ewing and Rodrigo Allen. Estimating population parameters using the structured serial coalescent with Bayesian MCMC inference when some demes are hidden. *Evolutionary Bioinformatics*, 2, 2006.

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

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

J Hey and R Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of drosophila pseudoobscura and d. persimilis. *Genetics*, 167:747–760, 2004.

J Hey and R Nielsen. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *PNAS*, 104:2785–2790, 2007.

Jody Hey, Yujin Chung, and Arun Sethuraman. On the occurrence of false positives in tests of migration under an isolation-with-migration model. *Molecular Ecology*, 24(20):5078–5083, 2015. ISSN 1365-294X. doi: 10.1111/mec.13381. URL `http://dx.doi.org/10.1111/mec.13381`.

Richard R Hudson et al. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology*, 7(1): 44, 1990.

Graham Jones. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of Mathematical Biology*, 2016. doi: 10.1007/s00285-016-1034-0. URL `http://link.springer.com/article/10.1007/s00285-016-1034-0`.

J.F.C. Kingman. The coalescent. *Stochastic processes and their applications*, 13(3):235–248, 1982.

A D Leaché, R B Harris, B Rannala, and Z Yang. The influence of gene flow on species tree estimation: A simulation study. *Systematic Biology*, 63(1):17–30, 2014.

Huw A. Ogilvie, Remco R. Bouckaert, and Alexei J. Drummond. Starbeast2 brings faster species tree inference and accurate estimates of substitution rates. *Mol Biol Evol*, 2017. doi: doi: 10.1093/molbev/msx126.

Michal Palczewski and Peter Beerli. A continuous method for gene flow. *Genetics*, 194(3):687–696, 2013. ISSN 0016-6731. doi: 10.1534/genetics.113.150904. URL `http://www.genetics.org/content/194/3/687`.

Claudia Solís-Lemus and Cécile Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, 12(3):1–21, 03 2016. doi: 10.1371/journal.pgen.1005896. URL `https://doi.org/10.1371/journal.pgen.1005896`.

Vitor C Sousa, Aude Grelaud, and Jody Hey. On the non-identifiability of migration time estimates in isolation with migration models. *Molecular ecology*, 20(19):3956, 2011.

Yuan Tian and Laura S Kubatko. Distribution of coalescent histories under the coalescent model with gene flow. *Molecular Phylogenetics and Evolution*, 105:177–192, 2016.