Divergence estimation in the presence of ILS and migration

Graham Jones

2017-04-17, May 9, 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 here to refer to gene flow between species (usually introgression but not restricted to that). A migration event occurs when an allele has 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 an earlier time than the 'source' branch. This is because coalesences are easier to model this way, and is the same convention as 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. The method of Dalquen et al. (2016) can estimate the species tree, but is currently restricted to at most 3 species and 3 sequences per locus, although it can handle very large numbers of loci.

Here we make two simplifying assumptions. Firstly, we use a simpler model for migration than IMa2 or Dalquen et al. (2016). The aim is to have a prior density for a gene tree, given the species tree, the population parameters, and the way in which the gene tree is embedded, which is straightforward to compute. Secondly, we approximate 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.

We call a pair of child branches of a gene tree node a **branch-pair**.

2 The prior density for an embedded gene tree

Suppose G_j are the parameters specifying the topology and node heights for the jth locus and S is similar parameters for the species tree. Suppose E_j are parameters which specify how the jth gene tree is to be embedded into the species tree.

The usual coalescent density for a gene trees $G = (G_1, G_2, \dots, G_J)$ given the species tree S and population size parameters Θ (Jones, 2016) can be extended to the case where there migrations between species tree branches. Each species tree branch b is divided by k_{jb} events (coalescences, migrations in or out) into $k_{jb} + 1$ intervals c_{jbi} $(0 \le i \le k_{jb})$, in which there are n_{jbi} gene tree lineages. This function (which we do not claim is a normalised density) is given by

$$\begin{aligned} f_{coal}(G; S, \Theta, E) &= \prod_j \prod_b (p_j \theta_b)^{-k_{jb}} \exp\Big(-\sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jbi}}{2} p_j \theta_b^{-1} \Big) \\ &= \prod_b r_b \theta_b^{-q_b} \exp\big(-\gamma_b \theta_b^{-1} \big) \end{aligned}$$

where

$$q_b = \sum_j k_{jb}, \qquad r_b = \prod_j p_j^{-k_{jb}}, \qquad \text{and} \ \ \gamma_b = \sum_j \sum_{i=0}^{k_{jb}} c_{jbi} \binom{n_{jbi}}{2} p_j^{-1}. \tag{1}$$

Now we turn to the probability of the embedding. For each branch pair i in any of the gene trees, set

$$p_{emb}(i; \nu) = \begin{cases} \nu & \text{if } i \text{ contains a migration} \\ 1 - \nu & \text{if } i \text{ soes not contain a migration} \end{cases}$$

Then we set

$$f_{emb}(G; S, \nu, E) = \prod_i p_{emb}(i; \nu)$$

Finally, we assume that the prior density for the embedded gene trees is given by

$$f(G|S, \Theta, \nu, E) \propto f_{emb}(G; S, \nu, E) f_{coal}(G; S, \Theta, E).$$

The proportional sign is present because f does not integrate to one. There is an interaction between the value of ν, the number of species, the population sizes, and the number of loci. In order to understand the properties of this prior, it is necessary to sample from it.

3 How the gene tree is embedded

This section describes the embedding parameters E_j, and how they are used to embed the gene trees. In principle, they could specify any possible embedding, but here we restrict the embeddings by ignoring ones which are unlikely when ν is small enough. The hope is that we will still explore a region of parameter space which includes most of the probability content. In order to assess how small ν needs to be for this to be a good approximation requires testing on simulations.

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)

The 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 1. The parameter ξ_{ji} determines where along a branch-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 branch-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.

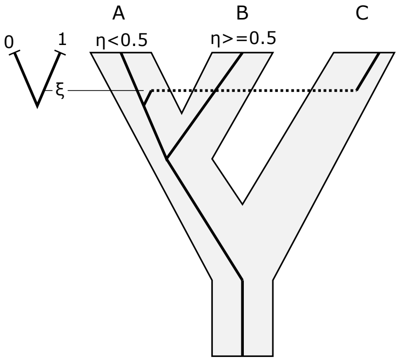

Figure 1: This shows ξ_{ji} and η_{ji} for a particular gene tree and the root node (with subscripts dropped). There is a migration from C into A. The parameter ξ determines how far along the branch-pair the migration occurs, and η determines whether the destination branch is A or B.

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 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, 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.

The prior $\Pr(E_j)$ for E_j is are iid uniform distributions on $[0,1]$ for each ξ_{ji} and η_{ji}. There is a hyper-prior for the migration intensity ν.

3.1 Implementation notes

In BEAST2, the node numbers are changed by operators and even by writing out a tree to a log file. The tree can be rotated at a node, so that the first/second or left/right children of some nodes are swapped. I use my own indices using an ordering based on an ordering of the tip labels.

DENIM uses the standard tree operators implemented in BEAST2. For a standard multispecies coalescent analysis, operators which change the species tree and the gene trees in a coordinated way are beneficial (Jones 2015, Olgilvie, Yna+Rannala). Here, any gene tree are compatible with any species tree, and these co-ordinated operators cannot be as-is. More work needed...

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.

3.2 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 2 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 2c). 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.

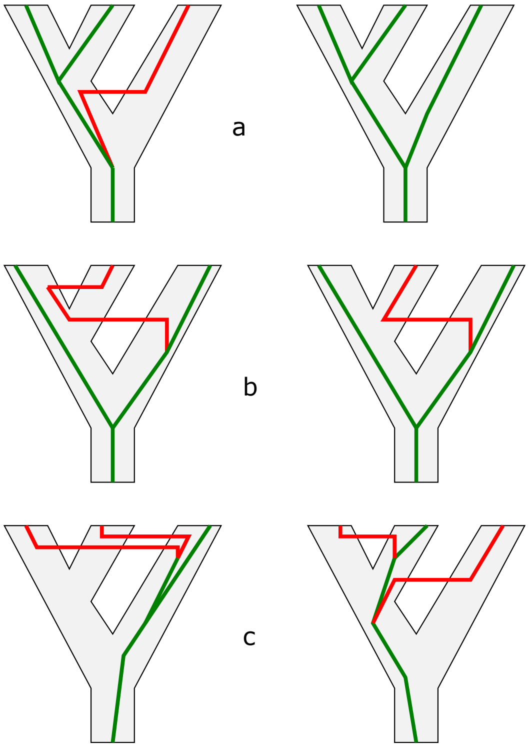

Figure 2: Some examples of embeddings for three gene trees in a, b, c. Branches which introgress are in red. On the left are embeddings that are ignored. On the right is an embedding which is considered.

4 Some results with prior only

The java code sets $1 - \nu$ to the following value, where `migIntensity=5.0` in the XML.

```
Math.exp(-0.015*migIntensity / sTree.getLeafNodeCount());
```

Thus ν is approximately 5*0.015/2 = 0.0375 for 2 species, and 5*0.015/6 = 0.0125 for 6 species. The following tables are made using `lookat-prior-logfiles.R`.

```
#        nspp nindivs nloci   ils nseqs sHeight ratio    ESS mean.migs migs.pbp
# [1,]      2       1     3 5e-02     6 0.00506  1.01 1423.4   0.00166  0.00166*
# [2,]      2       1     3 5e-03     6 0.00548  1.10 1325.6   0.02268  0.02268
# [3,]      2       1     3 5e-04     6 0.00596  1.19  726.7   0.07990  0.07990
# [4,]      2       1     9 5e-02    18 0.00505  1.01 1336.2   0.00298  0.00298*
# [5,]      2       1     9 5e-03    18 0.00607  1.21  606.3   0.02497  0.02497*
# [6,]      2       1     9 5e-04    18 0.00647  1.29  220.3   0.08614  0.08614

# [7,]      2       3     3 5e-02    18 0.00505  1.01 1326.0   0.02122  0.00424*
# [8,]      2       3     3 5e-03    18 0.00550  1.10  884.9   0.10362  0.02072
# [9,]      2       3     3 5e-04    18 0.00601  1.20  468.5   0.24449  0.04890
# [10,]     2       3     9 5e-02    54 0.00697  1.39  147.2   0.26183  0.05237
# [11,]     2       3     9 5e-03    54 0.00732  1.46  149.1   0.13372  0.02674
# [12,]     2       3     9 5e-04    54 0.00697  1.39  147.2   0.26183  0.05237

# [13,]     2       9     3 5e-02    54 0.00649  1.30  259.2   0.17853  0.01050*
# [14,]     2       9     3 5e-03    54 0.00705  1.41  273.4   0.52206  0.03071
# [15,]     2       9     3 5e-04    54 0.00592  1.18  106.0   0.67686  0.03982
# [16,]     2       9     9 5e-02   162 0.01140  2.28**264.6   0.27833  0.01637*
# [17,]     2       9     9 5e-03   162 0.01041  2.08**332.9   0.58711  0.03454
# [18,]     2       9     9 5e-04   162 0.00785  1.57* 126.2   0.73928  0.04349


         nspp nindivs nloci   ils nseqs sHeight ratio    ESS mean.migs  migs.pbp
# [19,]     6       1     3 5e-02    18 0.01456  1.00 1267.4   0.02538  0.00508*
# [20,]     6       1     3 5e-03    18 0.01503  1.04 1015.6   0.13754  0.02751
# [21,]     6       1     3 5e-04    18 0.01547  1.07  522.8   0.43508  0.08702
# [22,]     6       1     9 5e-02    54 0.01485  1.02  316.9   0.02566  0.00513*
# [23,]     6       1     9 5e-03    54 0.01573  1.08  155.8   0.16244  0.03249
# [24,]     6       1     9 5e-04    54 0.01663  1.15  205.4   0.45117  0.09023

# [25,]     6       3     3 5e-02    54 0.01606  1.11  380.8   0.19580  0.01152*
# [26,]     6       3     3 5e-03    54 0.01604  1.11  278.7   0.66708  0.03924
# [27,]     6       3     3 5e-04    54 0.01541  1.06  146.4   1.18498  0.06970
# [28,]     6       3     9 5e-02   162 0.01725  1.19  345.1   0.21577  0.01269*
# [29,]     6       3     9 5e-03   162 0.01939  1.34  415.7   0.77715  0.04571
# [30,]     6       3     9 5e-04   162 0.01947  1.34  211.1   1.27597  0.07506

# [31,]     6       9     3 5e-02   162 0.01850  1.28  339.2   1.29297  0.02440*
# [32,]    6       9     3 5e-03   162 0.01784  1.23  618.7   2.77237  0.05231
# [33,]    6       9     3 5e-04   162 0.01542  1.06  230.8   3.46546  0.06539
# [34,]    6       9     9 5e-02   486 0.02904  2.00**111     1.91662  0.03616
# [35,]    6       9     9 5e-03   486 0.02294  1.58* 191     3.06706  0.05787
# [36,]    6       9     9 5e-04   486 0.01885  1.30  207     3.57334  0.06742
```

- nspp = number of species
- nindivs = number of individuals per species
- nloci = number of loci
- ils = popPriorScale, which controls amount of ILS
- nseqs = total number of sequence (nspp x nindivs x nloci)
- sHeight = estimated species tree height. Should be 0.01*(1/2) = 0.005 for 2 species, 0.01*(1/2+1/3+1/4+1/5+1/6) = 0.0145 for 6 species
- ratio = (estimated species tree height) / (true species tree height)
- ESS = ESS for species tree height.
- mean.migs = average number of migration per locus. (Average taken over all MCMCsamples and all loci.)
- migs.pbp = average number of migrations per gene tree branch-pair

ESS for posterior generally much higher than for species tree height. The species tree gets stretched in the prior by the interaction between migration and the multispecies coalescent model. The worst cases are starred. The number of migrations per *branch-pair* is roughly around $\nu = 0.05$. Large deviations from this observation are starred.

5 Simulated data

This uses some data from Leaché et al. (2014). For now, there are 4 species (A,B,C,D), and the migration patterns do not include ones where alleles cross species boundaries at time 0. Twenty replicates. Also a further twenty replicates for my own MCcoal control files which should be sampling from the same distribution were analysed, but only the results for coverage are reported.

5.1 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	19	20
Ancestral-IM		19	20	19	20
Paraphyly		20	19	4	4
No migration	19				
Sister-1-allele	20				
Sister-1-mig	20				
Nonsister-1-allele	20				
Nonsister-1-mig	19				

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. TODO: A previous prior had worse results for the Nonsister-1-mig case, needs checking.

With no migration in the model, ($\nu = 0$), the results for the Paraphyly case were 19,9,4,4, similar to *BEAST results from Leaché et al. (2014), where the coverage values were 0.92, 0.55, 0.17, 0.03 for 100 replicates (they would 'predict' 18/20, 11/20, 4/20, 1/20).

5.2 Branch scores

TODO: time-zero changes with no migration in the model.

Figures 3 and 4 show the posterior means of tree distances between the true tree and the trees sampled during the MCMC algorithm. 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 only dramatic difference is for the paraphyly migration pattern with the migration rate at 0.01. There is modest improvement for the paraphyly migration pattern with the migration rate at 0.001 and 0.1, and for the IM migration pattern with migration rate at 0.001 and 0.01. TODO time-zero comparison.

5.3 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 at a small rate. For migration between sister species the performance is mixed.

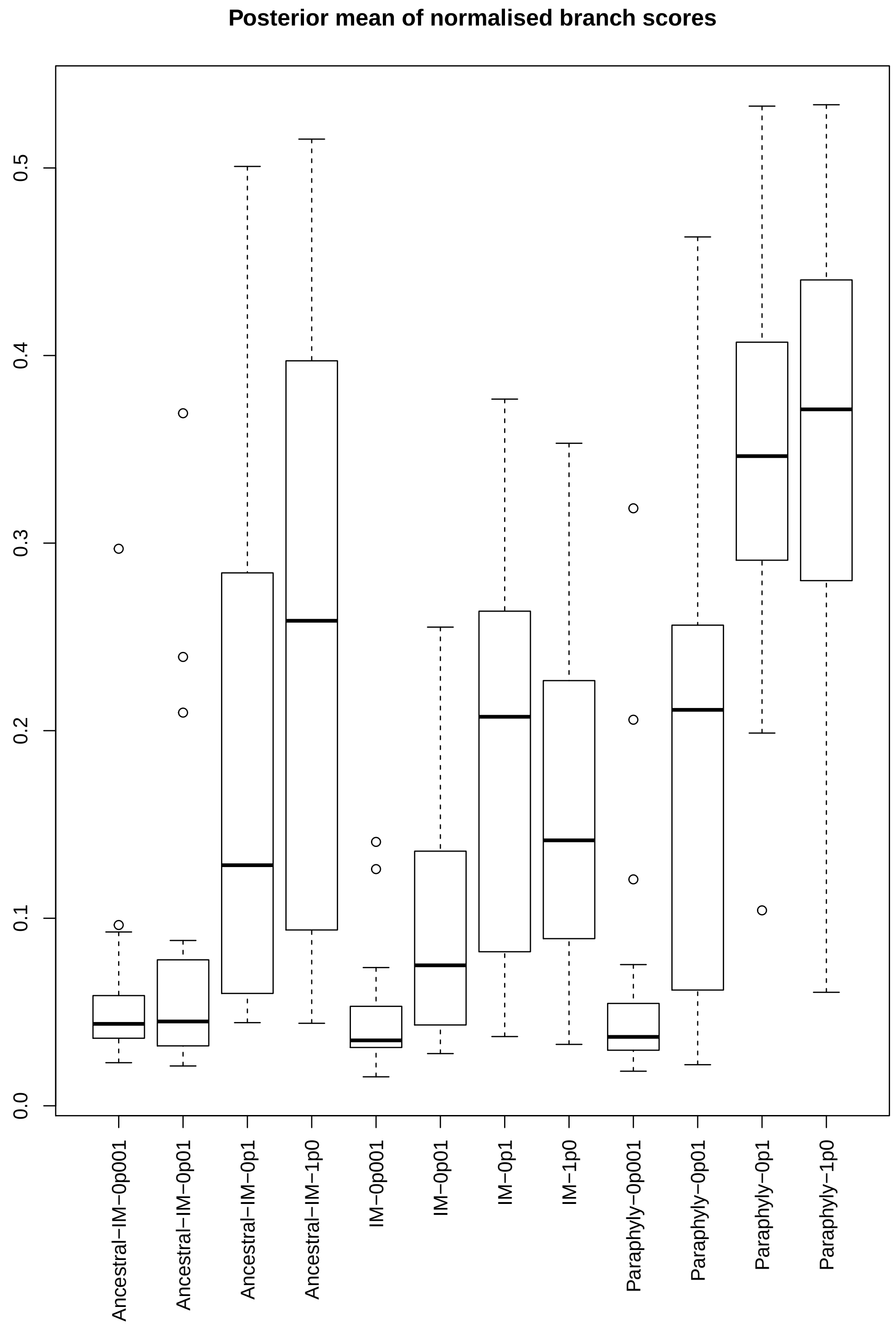

Figure 3: Branch scores with no migration in the model. (The prior on ν is exponential with mean 1e-50 which prevents any migration being considered. TODO add other cases.)

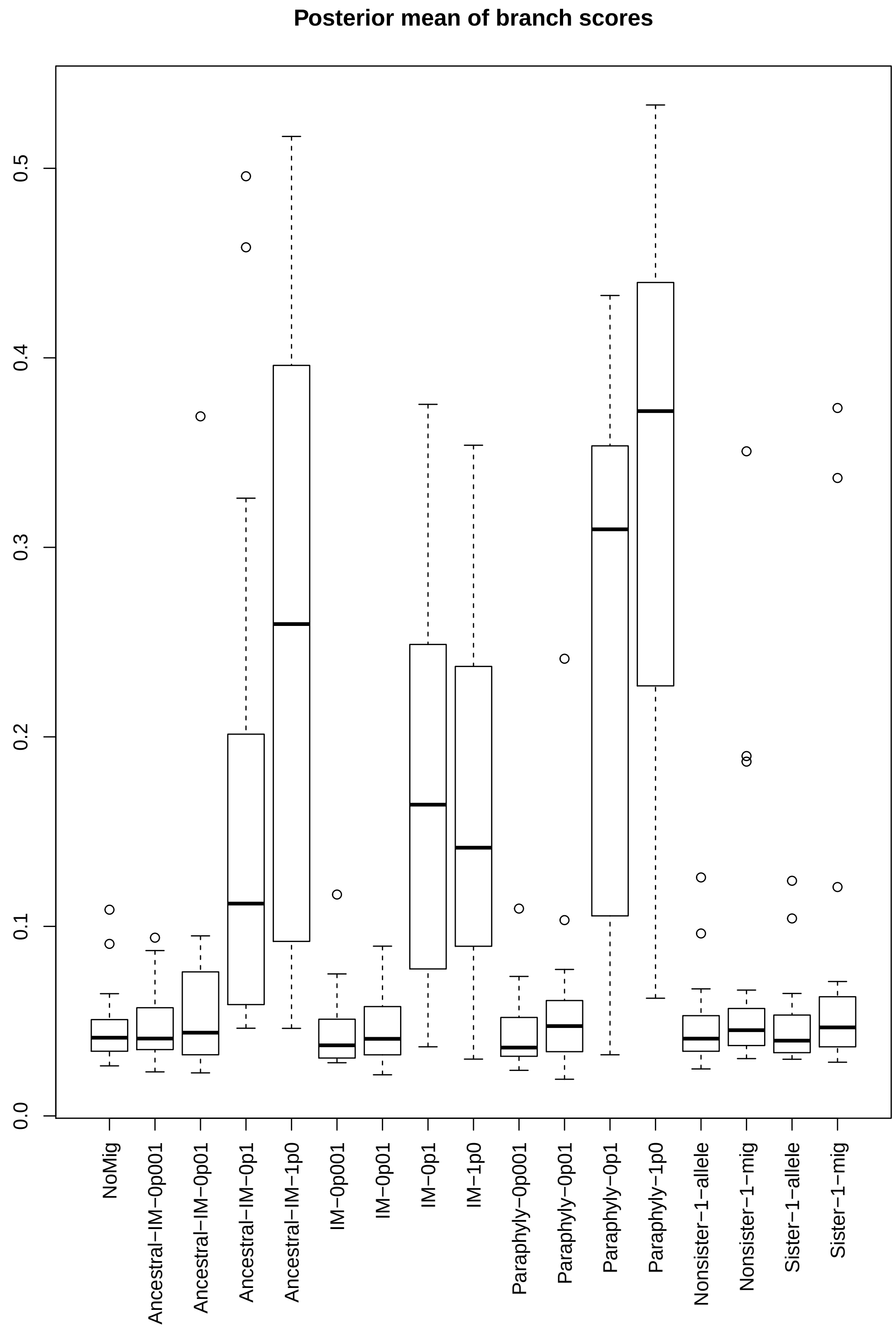

Figure 4: Branch scores with migration in the model. The prior on ν is exponential with mean 0.01.

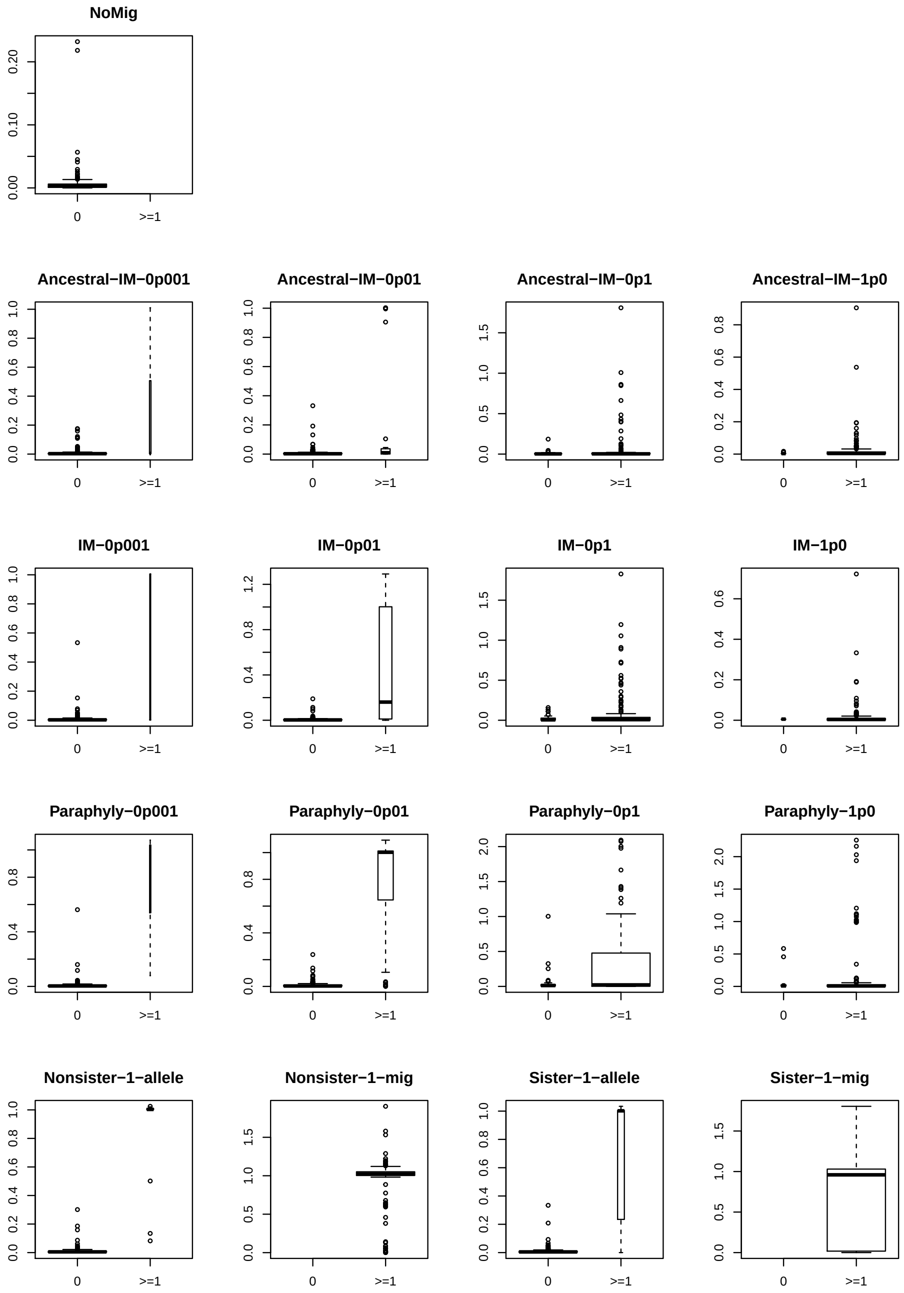

Figure 5: Posterior mean counts of migration, divided into two cases for each migration pattern and migration rate. Each boxplot is based on a total of 20 replicates and 10 loci. The label `0` means there is no migration at a locus in the simulated data, and the label `>=1` means there is some migration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

6 Discussion

The formula for f_{emb} was chosen in preference to various others that were tried because it seems to result in a more predictable prior.

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 usually(? often) be identified. Results for migration patterns only involving migration between sisters are slightly better than *BEAST or DENIM with $\nu = 0$. 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.

Clearly the model is a rather crude approximation to reality. The hope is that it is better to account for migration crudely than to ignore migration altogether.

Two compromises. A prior for gene tree density whose properties must be discovered. A partial sampling of the posterior. TODO.

I think that by including more information in E_j, it would be to consider every embedding which has a minimal number of migrations. However, this appears difficult to implement, and especially difficult to implement efficiently. It is not clear how much difference this issue makes to species tree estimation, since some minimal embeddings are always considered, and the ignored embeddings appear to be fairly rare, occurring only when two or more migrations are needed near to one another.

Extinction. If there are extinct species, migration can result in unusually deep coalescences. TODO

References

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.

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.

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`.

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.

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`.

More refs

Sousa VC, Grelaud A, Hey J. On the non-identifiability of migration time estimates in isolation with migration models. Molecular ecology. 2011;20(19):3956-3962.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.