Statistical phylogeography: Migrate-N and IMa

Nested clade analysis [11, 12, 13] represented the earliest attempt to develop a formal approach to using an estimate of phylogenetic relationships among haplotypes to infer something both about the biogeographic history of the populations in which they are contained and the evolutionary processes associated with the pattern of diversification implied by the phylogenetic relationships among haplotypes and their geographic distribution. The statistical parsimony part of NCA depends heavily on coalescent theory for calculating the "limits" of parsimony. As a result, NCA combines aspects of pure phylogenetic inference — parsimony — with aspects of pure population genetics — coalescent theory — to develop a set of inferences about the phylogeographic history of populations within species. NCA is now of primarily historical interest. So far as I am aware, no one uses it any more. Instead everyone uses methods based directly on coalescent theory or similar approaches. Before we get to that, though, I need to point out one complication that is taken for granted now that I first became aware of in the late 1980s when Pekka Pamilo and Mashatoshi Nei [10] pointed out that the phylogenetic relationships of a single gene might be different from those of the populations from which the samples were collected.

Gene trees *versus* population trees

There are several reasons why *gene trees* might not match *population trees*.

- It could simply be a problem of estimation. Given a particular set of gene sequences, we *estimate* a phylogenetic relationship among them. But our estimate could be wrong. In fact, given the astronomical number of different trees possible with 50 or 60 distinct sequences, every phylogenetic estimate is virtually certain to be wrong somewhere. We just don't know where. So a difference between our *estimate* of a gene tree could mean nothing more than that our gene tree estimate is wrong.

- There might have been a hybridization event in the past so that the phylogenetic history of the gene we're studying is different from that of the populations from which

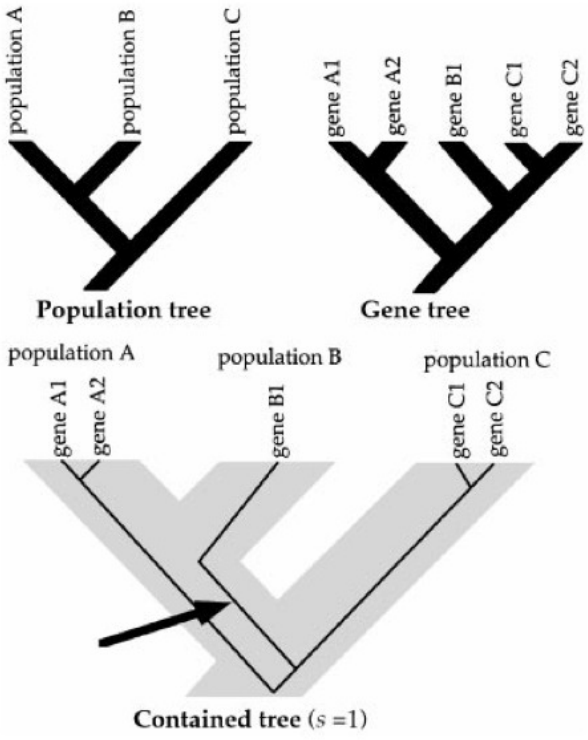

Figure 1: Discordance between gene and population trees as a result of ancestral polymorphism (from [7]).

we sampled. Hybridization is especially likely to have a large impact if the locus for which we have information is uniparentally inherited, e.g., mitochondrial or chloroplast DNA. A single hybridization event in the distant past in which the maternal parent was from a different population will give mtDNA or cpDNA a very different phylogeny than nuclear genes that underwent a lot of backcrossing after the hybridization event.

- If the ancestral population was polymorphic at the time the initial split occurred alleles that are more distantly related might, by chance, end up in the same descendant population (see Figure 1)

As Pamilo and Nei showed, it's possible to calculate the probability of discordance between the gene tree and the population tree using some basic ideas from coalescent theory. That leads to a further refinement, using coalescent theory directly to examine alternative biogeographic hypotheses.

Coalescent-based estimates of migration rate

Nearly 20 years ago Peter Beerli and Joe Felsenstein [3, 4] proposed a coalescent-based method to estimate migration rates among populations. As with other analytical methods we've encountered in this course, the details can get pretty hairy, but the basic idea is (relatively) simple.

Recall that in a single population we can describe the coalescent history of a sample without too much difficulty. Specifically, given a sample of k alleles in a diploid population with effective size N_e, the probability that the first coalescent event took place t generations ago is

$$P(t|k, N_e) = \left(\frac{k(k-1)}{4N_e}\right)\left(1 - \frac{k(k-1)}{4N_e}\right)^{t-1} \quad . \tag{1}$$

Now suppose that we have a sample of alleles from K different populations. To keep things (relatively) simple, we'll imagine that we have a sample of n alleles from every one of these populations and that every population has an effective size of N_e. In addition, we'll imagine that there is migration among populations, but again we'll keep it really simple. Specifically, we'll assume that the probability that a given allele in our sample from one population had its ancestor in a different population in the immediately preceding generation is m.[1] Under this simple scenario, we can again construct the coalescent history of our sample. How? Funny you should ask.

We start by using the same logic we used to construct equation (1). Specifically, we ask "What's the probability of an 'event' in the immediately preceding generation?" The complication is that there are two kinds of events possible:

1. a coalescent event and

2. a migration event.

As in our original development of the coalescent process, we'll assume that the population sizes are large enough that the probability of two coalescent events in a single time step is so small as to be negligible. In addition, we'll assume that the number of populations and the migration rates are small enough that the probability of more than one event of either type is so small as to be negligible. That means that all we have to do is to calculate the probability of either a coalescent event or a migration event and combine them to calculate the probability of an event. It turns out that it's easiest to calculate the probability that

[1] In other words, m is the backwards migration rate, the probability that a gene in one population came from another population in the preceding generation. This is the same migration rate we encountered weeks ago when we discussed the balance between drift and migration.

there *isn't* an event first and then to calculate the probability that there is an event as one minus that.

We already know that the probability of a coalescent event in population k, is

$$P_k(\text{coalescent}|n, N_e) = \frac{k(k-1)}{4N_e} \quad ,$$

so the probability that there is *not* a coalescent event in any of our K populations is

$$P(\text{no coalescent}|k, N_e, K) = \left(1 - \frac{k(k-1)}{4N_e}\right)^K \quad .$$

If m is the probability that there was a migration event in a particular population than the probability that there is *not* a migration event involving any of our kK alleles[2] is

$$P(\text{no migration}|k, m, K) = (1-m)^{kK} \quad .$$

So the probability that there *is* an event of some kind is

$$P(\text{event}|k, m, N_e, K) = 1 - P(\text{no coalescent}|k, N_e, K)P(\text{no migration}|k, m, K) \quad .$$

Now we can calculate the time back to the first event

$$P(\text{event at } t|k, m, N_e, K) = P(\text{event}|k, m, N_e, K)\,(1 - P(\text{event}|k, m, N_e, K))^{t-1} \quad .$$

We can then use Bayes theorem to calculate the probability that the event was a coalescence or a migration and the population or populations involved. Once we've done that, we have a new population configuration and we can start over. We continue until all of the alleles have coalesced into a single common ancestor, and then we have the complete coalescent history of our sample.[3] That's roughly the logic that Beerli and Felsenstein use to construct coalescent histories for a sample of alleles from a set of populations — except that they allow effective population sizes to differ among populations and they allow migration rates to differ among all pairs of populations. As if that weren't bad enough, now things start to get even more complicated.

There are lots of different coalescent histories possible for a sample consisting of n alleles from each of K different populations, even when we fix m and N_e. Worse yet, given any

[2] K populations each with k alleles

[3] This may not seem very simple, but just think about how complicated it would be if I allowed every population to have a different effective size and if I allowed each pair of populations to have different migration rates between them.

one coalescent history, there are a lot of different possible mutational histories possible. In short, there are a lot of different possible sample configurations consistent with a given set of migration rates and effective population size. Nonetheless, some combinations of m and N_e will make the data more likely than others. In other words, we can construct a likelihood for our data:

$$P(\text{data}|m, N_e) \propto f(n, m, N_e, K) \quad ,$$

where $f(n, m, N_e, K)$ is some very complicated function of the probabilities we derived above. In fact, the function is so complicated, we can't even write it down. Beerli and Felsenstein, being very clever people, figured out a way to simulate the likelihood, and `Migrate-n` `http://popgen.sc.fsu.edu/Migrate/Migrate-n.html` provides a (relatively) simple way that you can use your data to estimate m and N_e for a set of populations. In fact, `Migrate-N` will allow you to estimate pairwise migration rates among all populations in your sample, and since it can simulate a likelihood, if you put priors on the parameters you're interested in, i.e., m and N_e, you can get Bayesian estimates of those parameters rather than maximum likelihood estimates, including credible intervals around those estimates so that you have a good sense of how reliable your estimates are.[4]

There's one further complication I need to mention, and it involves a lie I just told you. `Migrate-N` can't give you estimates of m and N_e. Remember how every time we've dealt with drift and another process we always end up with things like $4N_em$, $4N_e\mu$, and the like. Well, the situation is no different here. What `Migrate-N` can actually estimate are the two parameters $4N_em$ and $\theta = 4N_e\mu$.[5] How did μ get in here when I only mentioned it in passing? Well, remember that I said that once the computer has constructed a coalescent history, it has to apply mutations to that history. Without mutation, all of the alleles in our sample would be identical to one another. Mutation is what what produces the diversity. So what we get from `Migrate-N` isn't the fraction of a population that's composed of migrants. Rather, we get an estimate of how much migration contributes to local population diversity relative to mutation. That's a pretty interesting estimate to have, but it may not be everything that we want.

There's a further complication to be aware of. Think about the simulation process I described. All of the alleles in our sample are descended from a single common ancestor. That means we are implicitly assuming that the set of populations we're studying have been around long enough and have been exchanging migrants with one another long enough that we've reached a drift-mutation-migration equilibrium. If we're dealing with a relatively small

[4]If you'd like to see a comparision of maximum likelihood and Bayesian approaches, Beerli [1] provides an excellent overview.

[5]Depending on the option you pick when you run `Migrate` you can either get θ and $4N_em$ or θ and $M = m/\mu$.

number of populations in a geographically limited area, that may not be an unreasonable assumption, but what if we're dealing with populations of crickets spread across all of the northern Rocky Mountains? And what if we haven't sampled all of the populations that exist?[6] In many circumstances, it may be more appropriate to imagine that populations diverged from one another at some time in the not too distant past, have exchanged genes since their divergence, but haven't had time to reach a drift-mutation-migration equilibrium. What do we do then?

Divergence and migration

Rasmus Nielsen and John Wakely [8] consider the simplest generalization of Beerli and Felsenstein [3, 4] you could imagine (Figure 2). They consider a situation in which you have samples from only two populations and you're interested in determining both how long ago the populations diverged from one another and how much gene exchange there has been between the populations since they diverged. As in `Migrate-N` mutation and migration rates are confounded with effective population size, and the relevant parameters become:

- θ_a, which is $4N_e\mu$ in the ancestral population.
- θ_1, which is $4N_e\mu$ in the first population.
- θ_2, which is $4N_e\mu$ in the second population.
- M_1, which is $2N_em_1$ in the first population, where m_1 is the fraction of the first population composed of migrants from the second population.
- M_2, which is $2N_em_2$ in the second population.
- T, which is the time since the populations diverged. Specifically, if there have been t units since the two populations diverged, $T = t/2N_1$, where N_1 is the effective size of the first population.

Given that set of parameters, you can probably imagine that you can calculate the likelihood of the data for a given set of parameters.[7] Once you can do that you can either

[6] Beerli [2] discusses the impact of "ghost" populations. He concludes that you have to be careful about which populations you sample, but that you don't necessarily need to sample every population. Read the paper for the details.

[7] As with `Migrate-N`, you can't calculate the likelihood explicitly, but you can approximate it numerically. See [8] for details.

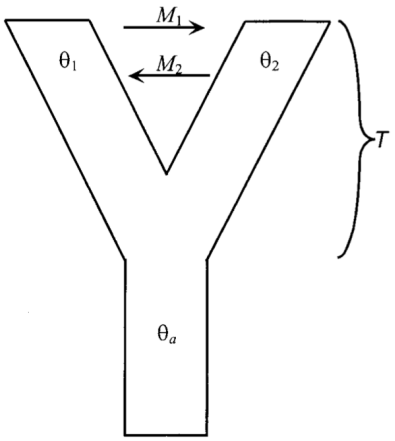

Figure 2: The simple model developed by Nielsen and Wakeley [8]. θ_a is $4N_e\mu$ in the ancestral population; θ_1 and θ_2 are $4N_e\mu$ in the descendant populations; M_1 and M_2 are $2N_em$, where m is the backward migration rate; and T is the time since divergence of the two populations.

obtain maximum-likelihood estimates of the parameters by maximizing the likelihood, or you can place prior distributions on the parameters and obtain Bayesian estimates from the posterior distribution. Either way, armed with estimates of θ_a, θ_1, θ_2, M_1, M_2, and T you can say something about:

1. the effective population sizes of the two populations relative to one another and relative to the ancestral population,

2. the relative frequency with which migrants enter each of the two populations from the other, and

3. the time at which the two populations diverged from one another.

Keep in mind, though, that the estimates of M_1 and M_2 confound local effective population sizes with migration rates. So if $M_1 > M_2$, for example, it does not mean that the fraction of migrants incorporated into population 1 exceeds the fraction incorporated into population 2. It means that the impact of migration has been felt more strongly in population 1 than in population 2.

An example

Orti et al. [9] report the results of phylogenetic analyses of mtDNA sequences from 25 populations of threespine stickleback, *Gasterosteus aculeatus*, in Europe, North America, and Japan. The data consist of sequences from a 747bp fragment of cytochrome *b*. Nielsen and Wakely [8] analyze these data using their approach. Their analyses show that "[a] model of moderate migration and very long divergence times is more compatible with the data than a model of short divergence times and low migration rates." By "very long divergence times" they mean $T > 4.5$, i.e., $t > 4.5N_1$. Focusing on populations in the western (population 1) and eastern Pacific (population 2), they find that the maximum likelihood estimate of M_1 is 0, indicating that there is little if any gene flow from the eastern Pacific (population 2) into the western Pacific (population 1). In contrast, the maximum likelihood estiamte of M_2 is about 0.5, indicating that one individual is incorporated into the eastern Pacific population from the western Pacific population every other generation. The maximum-likelihood estimates of θ_1 and θ_2 indicate that the effective size of the population eastern Pacific population is about 3.0 times greater than that of the western Pacific population.

Extending the approach to multiple populations

A little over five years ago, Jody Hey announced the release of `IMa2`.[8] implement this method. Building on work described in Hey and Nielsen [5, 6], `IMa2` allows you to estimate relative divergence times, relative effective population sizes, and relative pairwise migration rates for more than two populations at a time. That flexibility comes at a cost, of course. In particular, you have to specify the phylogenetic history of the populations before you begin the analysis.

References

[1] P Beerli. Comparison of Bayesian and maximum-likelihood estimation of population genetic parameters. *Bioinformatics*, 22:341–345, 2006.

[2] Peter Beerli. Effect of unsampled populations on the estimation of population sizes and migration rates between sampled populations, 2004.

[3] Peter Beerli and Joseph Felsenstein. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach, 1999.

[8] Available from `https://bio.cst.temple.edu/~hey/software/software.htm`.

[4] Peter Beerli and Joseph Felsenstein. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach, 2001.

[5] Jody Hey and Rasmus Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*, 2004.

[6] Jody Hey and Rasmus Nielsen. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8):2785–2790, 2007.

[7] L Knowles. Did the Pleistocene glaciations promote divergence? Tests of explicit refugial models in montane grasshopprers, 2001.

[8] Rasmus Nielsen and J Wakeley. Distinguishing migration from isolation: a Markov chain Monte Carlo approach, 2001.

[9] Guillermo Orti, Michael A Bell, Thomas E Reimchen, and Axel Meyer. Global survey of mitochondrial DNA sequences in the threespine stickleback: evidence for recent migrations. *Evolution*, 48(3):608–622, 1994.

[10] P Pamilo and M Nei. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5):568–583, 1988.

[11] Alan R Templeton. Statistical phylogeography: methods of evaluating and minimizing inference errors, 2004.

[12] Alan R Templeton. Statistical hypothesis testing in intraspecific phylogeography: nested clade phylogeographical analysis vs. approximate Bayesian computation, 2009.

[13] Alan R Templeton, Eric Routman, and Christopher A Phillips. Separating population structure from population history: a cladistic analysis of the geographical distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics*, 140(2):767–782, 1995.

Creative Commons License