Learning about modes of speciation by computational approaches: good and bad news

Celine Becquet and Molly Przeworski
Dept. of Human Genetics, University of Chicago, Chicago IL USA

Background
Enduring debate in evolutionary biology centers around the question of whether the early stages of speciation can occur in the presence of gene flow.

Allopatric speciation
• Early stage of speciation occurs in the absence of gene flow, when an exogenous barrier isolates the populations.
• Divergence occurs homogeneously across the genome, as the populations accumulate differences through either genetic drift or local adaptation.

Parapatric & sympatric speciation
• Species start diverging while hybridizing and exchanging migrants (i.e., they occupy contiguous areas, or the same area).
• Divergence occurs in a number of stages:
  1. Natural selection
  2. Population subdivision
  3. Recombination

Recombining loci under I-M models but with assumptions of the model violated. The effect parameter estimates of the I-M model.

Nielsen (2004) and more generally, investigated how violations of model assumptions affect parameter estimates of the I-M model.

Recently, computational approaches have been developed to test the predictions of simple models of divergence (usually an I-M model). Application of these to a variety of parapatry, barrier isolates the populations.

Divergence occurs in a number of stages: 1. Natural selection 2. Population subdivision 3. Recombination in some constant gene flow is prevented throughout the genome. These modes are sometimes approximated by the isolation model. I-M.

Modeling as well as from empirical studies suggest that the early stages of speciation may often occur in presence of some gene flow. However, the parapatric model of speciation remains controversial, in part because of the difficulty of distinguishing parapatry from allopatry followed by secondary contact.

Methods
Simulated data. We simulated data sets of 20 independently-evolving 1 Mb non-recombining loci under I-M models but with assumptions of the model violated. The parameter values were such that \( T = 3.2 N_e \). We generated 10 data sets for each of the combinations of parameters: \( f = 0.25, 0.50 \) and 0.75 (where \( f \times 2 \times 10^7 \) generations), and \( N = 10, 50, 100 \) and 500.

a) Model of isolation from a structured ancestral population

b) Model of isolation followed by secondary contact

c) Model of isolation with migration at an early stage.

Data with a modeled RI locus.

We assumed that a sample of 20 independently-evolving loci contains one locus closely linked to a RI factor. Specifically, polymorphism data at the loci were simulated under a model of I-M with one outlier that has experienced either no gene flow since the split (model I), or positive directional selection in one population. The data were simulated using a program that we recently developed (MIMAR) as well as the program I-M of Hey and Nielsen (2004) and more generally, investigated how violations of model assumptions affect parameter estimates of the I-M model.

Estimation approaches. We analyzed the simulated data with MIMAR and I-M.

• Both programs use a Markov Chain Monte Carlo approach and rely on multi-locus polymorphism data to estimate the parameters of the I-M model.
• I-M relies on the full polymorphism data, but assumes no intro-locus recombination, a limitation that can lead to biased estimates.

• MIMAR estimates parameters of the I-M model from recombining loci. In contrast to I-M, it relies on summaries of the polymorphism data at each of multiple independently-evolving loci – specifically, four summary statistics that are sensitive to the parameters of the I-M model:
  • the number of derived polymorphisms unique to the sample from populations \( 1 \) and \( 2 \) (and \( 3 \));
  • the number of fixed derived alleles between the two samples (1, 2);
  • the number of fixed derived alleles in either sample (1, 2) assuming a known ancestral state.

• In simulations, MIMAR performs comparably to I-M for medium size data sets, providing unbiased estimates of the parameters and reasonable power to detect gene flow (Becquet and Przeworski 2007).

Results
Allopatric divergence from a structured ancestral population
Applications of MIMAR and I-M to real data often yield an estimate of the ancestral effective population size larger than either of the descendant populations. Could such results reflect a case of parapatric speciation? We investigated this possibility by estimating the parameters of the I-M model from data simulated under models of isolation from a structured ancestral population (model c).

• I-M tends to provide large estimates of \( N_e \) when there is structure in the ancestral population.

• Nonetheless, the results of I-M would usually be interpreted correctly as allopatric speciation (i.e., the estimated gene flow rate is not biased).

• MIMAR does seem to provide biased estimates of \( N_e \).

• However, the results would often be interpreted incorrectly as rejecting a model of allopatric speciation (i.e., the estimated gene flow rate tends to be biased upwards).

Parapatry with gene flow at an early stage.

We investigated the effect of a change in gene flow rates over time on the parameter estimates (model c).

• Estimates of \( N_e \) are qualitatively similar than for model a, thus, large estimates of \( N_e \) provided by I-M on real data may reflect not constant gene flow since the split.

• Moreover, I-M tends not to detect gene flow, which could be interpreted incorrectly as consistent with diploidy, speciation.

• Both methods tends to underestimate \( T \) in this case.

Detecting loci linked to a RI factor or a target of local selection.
To date, only a few genes involved in RI between species have been characterized, through time-consuming molecular approaches. Could computational approaches such as MIMAR be used to help identify candidate loci for further investigation? To assess this, we simulated two series of data sets in which one locus is assumed to be linked to a RI factor, using models 1 and 2, and analyzed them using MIMAR.

• MIMAR does not seem greatly affected by the inclusion of a locus with a different history.

• A simple locus-specific goodness-of-fit tests may help to detect the locus of interest.
• We obtained the posterior predictive distribution of statistics for each single locus in the data set given the data simulated by MIMAR (see Table below).

Conclusion
We find that when one of many assumptions of the I-M model are violated, the methods tend to yield upwardly biased estimates of gene flow, potentially lending spurious support for parapatric speciation. To some extent, this problem can be avoided by carefully testing the fit of estimated models to the data. When a parapatric model is appropriate, we propose a test that can help to pinpoint candidate loci involved in RI.

References
Becquet and Przeworski 2007 Genome Res 17:1505-1519
Hey and Nielsen 2007 Genetics 175:713-743
Hey 2005 Curr Opin Genet Dev 15:992-996