The confounding effects of incomplete lineage sorting and selective sweeps on reconstructing species trees: an empirical multilocus study of firefinches (*Lagonosticta* spp.)

Heather Shull, Boston University

Introduction – A major issue in phylogenetic reconstruction is the assumption that gene trees adequately represent species' history. Gene trees from unlinked loci, each showing the genealogy of a set of DNA sequences, and species trees, representing the evolutionary history of speciation, are not equivalent in theory and are often incongruent in practice (Avise, 1989). Allelic polymorphism in an ancestral population renders gene trees for newly divergent species polyphyletic, and only new mutations and the stochastic loss of ancestral allelic lineages by genetic drift, termed "lineage sorting", will lead to genetic monophyly of sister species. The time it takes for this process of lineage sorting to occur is related to population sizes and the level of polymorphism in the ancestral and descendant species (Pamilo & Nei, 1988). If subsequent speciation events occur before lineage sorting is complete, species may retain ancestral allelic lineages that do not accurately represent the history of divergence in the group.

The expected relationship between the rate of lineage sorting and population size does not necessarily hold true for loci linked to regions under selection. Selective sweeps, in which a novel adaptive mutation sweeps to fixation, fix the haplotypes of closely linked regions and rapidly reduce genetic diversity (Barton, 2000). Larger populations should experience this more often because of the increased opportunity for mutation, potentially decoupling the relationship between genetic diversity and population size. Furthermore, low levels of introgressive hybridization could introduce new, advantageous alleles into a population, the fixation of which would result in misleading phylogenetic inferences from these loci and misinformation about the level of genetic divergence between species. In a recent study, nuclear but not mitochondrial genetic diversity correlated with relative population size, indicating the frequent occurrence of selective sweeps in the non-recombining mitochondrial genome (Bazin et al., 2006). The above study, however, made comparisons among highly divergent taxonomic groups (mammals vs. insects, for example), such that differences in mating systems and relative mutation rates might dramatically influence expected relationships between genetic diversity and population size.

Objectives – The objectives of this study are 1) to empirically analyze the relationship between historical demography and the prevalence of incomplete lineage sorting across an entire clade of estrildid finches, the African firefinches (*Lagonosticta* spp.); and 2) to test for the occurrence of selective sweeps in the mitochondrial genome of this clade. *Lagonosticta* comprises ten closely related species that show dramatic differences in range size (for example, *L. sanguinodorsalis* is endemic to a single plateau in Nigeria, while *L. senegala* is widespread across most of sub-Saharan Africa) and that show reciprocal monophyly at mitochondrial loci (Sorenson et al., 2004). I will sequence multiple nuclear loci to test for predicted effects of population size and speciation intervals (i.e., length of internodes) on the extent of gene tree incongruence and genetic diversity. As a test for selective sweeps in the mitochondrial genome, the empirical mtDNA data will be compared to expectations derived from simulations using estimates of demographic parameters and divergence times from nuclear loci.

Methods – I have compiled a set of 65 firefinch tissue and feather samples from recent field work and the University of Michigan tissue collection that captures all 10 species and much of the intraspecific taxonomic and geographic variation, critical because widespread species may not freely interbreed across the entire geographic range. Using standard methods, two mitochondrial genes (ND2 and ND6) and 20 anonymous nuclear loci will be sequenced for each sample, primers for which were developed from a genomic library of indigobirds, a closely related group of finches.

As an approximate measure of population size, species will be categorized as widespread (n = 3), intermediate (n = 4), or restricted range (n = 3). Examining a closely related set of species with extreme differences in range size but which share similar mating patterns (i.e., monogamy with biparental care) will help isolate the effects of population size per se while reducing bias due to interspecific variation in effective population size relative to census size.

Nuclear haplotypes will be resolved using the program PHASE (Stephens et al., 2001) and supplemented with bacterial cloning as necessary, and a hypothesis of phylogenetic relationships for the alleles at each locus will be estimated using PAUP* (Swofford, 2002). Trees will be rooted with sequences from *Clytospiza monteiri*, the monotypic sister group of firefinches (Sorenson et al., 2004). An overall estimate of species relationships and relative divergence times will be inferred from the combined nuclear sequence data using developing methods that incorporate the lineage sorting process of gene trees into estimating species histories (Maddison, 1997; Liu & Pearl, 2006). Most or all of the anonymous nuclear loci are presumably neutral and not closely linked to regions under selection, resulting in an unbiased inference of evolutionary relationships.

I will examine the frequency of incomplete lineage sorting by testing for a correlation between the level of gene tree incongruence and relative population size. Incongruence will be calculated as the percentage of gene trees that are inconsistent with monophyly for each species and higher level node in the tree. As time since speciation will also affect the extent of gene tree incongruence, I will also consider relative branch lengths in the concatenated phylogeny (as a measure of time). The basic prediction is that species and species groups with larger ranges and less time since speciation (or between speciation events) will show monophyly less frequently than those with more restricted ranges and/or longer times since speciation.

To test for the occurrence of selective sweeps, I will compare the observed level of monophyly and intraspecific genetic diversity in the mtDNA gene tree with the expected distribution of mtDNA gene trees simulated using parameters estimated from the inferred speciational history. First, I will estimate ancestral and descendant population sizes and relative divergence times using the nuclear data and the Bayesian method implemented in the program MCMCcoal (Rannala & Yang, 2003). These parameter values will then be used to simulate the distribution of gene trees expected from mtDNA (Degnan & Salter, 2005), given the ¼ effective population size compared to nuclear markers. If non-monophyly or deeper intraspecific coalescent times are expected based on these demographic parameters compared to the observed mitochondrial gene tree, that would suggest that selective sweeps have reduced diversity and increased the rate of lineage sorting at mitochondrial loci. This would also eliminate the correlation between population size and genetic diversity, so I will also compare nuclear and mitochondrial estimates of population size and determine if they correlate with relative population sizes estimated from range distributions.

Significance – Incomplete lineage sorting and selective sweeps are often invoked as *ad hoc* explanations for unexpected phylogenetic relationships, but as multiple, unlinked nuclear sequence data become the new standard in phylogenetic reconstruction, these issues will become of primary importance. The incongruence of nuclear gene trees has been predicted based on coalescent simulations (Maddison & Knowles, 2006) and documented in groups of 3-4 species using a single individual per species (Edwards and Jennings, 2005; Pollard et al., 2006), but incomplete lineage sorting has the potential to extend beyond single speciation events among closely related species. This study will therefore thoroughly explore the effect of population size and speciation time on the extent of incomplete lineage sorting at nuclear and mitochondrial markers in a much larger empirical context than has been studied to date.

The tendency of mitochondrial data to show genetic monophyly (due to its faster rate of lineage sorting) has made it a useful marker for phylogenetic and population level studies. This study will examine whether this increased monophyly is also influenced by the occurrence of selective sweeps, by comparing the gene tree expectations derived from demographic parameters with the estimated mitochondrial gene tree. The hypothesis that selective sweeps have increased the rate of lineage sorting is supported by preliminary data; the mitochondrial gene tree shows monophyletic species with strongly supported relationships but nuclear gene trees show a high level of incongruence. Documenting the occurrence and frequency of selective sweeps is critical to understanding the evolution of the mitochondrial genome, the effect of divergence and introgression on resulting patterns of genetic diversity, and the processes by which species adapt and respond to natural selection (Barton, 2000).

Finally, this large empirical dataset will be valuable for comparing developing methods of phylogenetic reconstruction that incorporate the lineage sorting process through the use of coalescent models. This study is therefore well-poised to address key issues about which most systematists are aware but for which there is still a paucity of published empirical research that can inform future studies dealing with multiple nuclear loci.

Schedule – This project constitutes one aspect of my dissertation, which focuses on the interaction of demography and natural selection in shaping genetic diversity in brood parasitic finches (*Vidua* spp.) and their firefinch hosts. This is my third year in the PhD program at Boston University, but I have developed this project over the past 9 months as I have encountered some of the difficulties in working with nuclear data. To date, I have sequenced mtDNA and nine nuclear loci for 29 of 65 individual specimens. Sequencing of the remaining loci and samples will continue through 2007, be analyzed in early 2008, and submitted for publication later that year.

Literature Cited

Avise JC. 1989. Gene trees and organismal histories: a phylogenetic approach to population biology. Evolution 43: 1192-1208.

Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312: 570-572.

Barton, NH. 2000. Genetic Hitchhiking. Phil Trans R Soc Lond B Biol Sci 355: 1553-1562.

Degnan JH, Salter LA. 2005. Gene tree distributions under the coalescent process. Evolution 59: 24-37.

Jennings WB, Edwards SV. 2005. Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. Evolution 59: 2033-2047.

Liu L, Pearl DK. 2006. Reconstructing posterior distributions of a species phylogeny using estimated gene tree distributions. Mathematical Biosciences Institute Technical Report #53. The Ohio State University, Columbus, Ohio.

Maddison WP. 1997. Gene trees in species trees. Syst Biol 46: 523-536.

Maddison WP, Knowles LL. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst Biol 55: 21-30.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. Mol Biol Evol 5: 568-583.

PollardDA, Venky NI, Moses AM, Eisen MB. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: Evidence for incomplete lineage sorting. PLOS Genetics 2: 1634-1647.

Rannala B, Yang ZH. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645-1656.

Sorenson MD, Balakrishnan CN, Payne RB. 2004. Clade-limited colonization in brood parasitic finches (*Vidua* spp.). Syst Biol 53:140-153.

Stephens M, Smith MJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. Am J Human Genetics 68: 978-989.

Swofford DL. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, MA

Budget Justification:

This research will require 65 DNA extractions and approximately 1430 PCR and sequencing reactions (loci are ~300 bp and therefore require only a single strand to be sequenced). Twenty-nine extractions and 320 PCR and sequencing reactions have already been completed. PCR and sequencing is done in high volume in the lab and costs are therefore reduced as much as possible by purchasing in bulk. Cloning will be required for those sequences with unresolvable haplotypes, but I will pool samples from various loci to reduce the number of cloning reactions needed and expect to need approximately 30 reactions, which will add an additional 300 PCR and sequencing reactions.

Itemized costs—

DNA Extractions:

$80—Qiagen DNA Extraction kit ($112/50 reactions)

Cloning:

$307—TOPO TA cloning kit ($410/20 reactions, ½ reaction size)

PCR costs:

$99—DNA polymerase AmpliTaq Gold ($1,410/5000 units, ¼ reaction size)

$225—Exo/SAP for PCR clean-up ($0.16/sample)

Sequencing costs:

$553—ABI big dye sequencing kit ($31,400/5000 rxn kit, 1/16th reaction size)

$381—Sequence analysis fees ($13/48 samples)

Total: $1,645