Spring 4-14-2011

Multilocus and parametric analyses of the evolutionary history of the Amazonian peacock cichlids, the genus Cichla (Teleostei: Cichlidae)

Stuart Willis
University of Nebraska-Lincoln, swillis4@gmail.com

MULTILOCUS AND PARAMETRIC ANALYSES OF THE EVOLUTIONARY HISTORY OF THE AMAZONIAN PEACOCK CICHLIDS, GENUS *CICHLA* (TELEOSTEI: CICHLIDAE)

by

Stuart Clayton Willis

A DISSERTATION

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Doctor of Philosophy

Major: Biological Sciences

Under the Supervision of Professors Diana Pilson and Guillermo Ortí

Lincoln, Nebraska

April, 2011

MULTILOCUS AND PARAMETRIC ANALYSES OF THE EVOLUTIONARY HISTORY OF THE AMAZONIAN PEACOCK CICHLIDS, GENUS *CICHLA* (TELEOSTEI: CICHLIDAE)

Stuart C. Willis, Ph.D.

University of Nebraska, 2011

Advisors: Diana Pilson and Guillermo Ortí

Accurate knowledge of species boundaries and species phylogeny are fundamental to testing hypotheses of recent evolutionary processes, but the estimation of these partitions is challenging due both to inherent confusion about what is being estimated as well as the data available to estimate them. Using multilocus data from mtDNA, microsatellites, and nuclear locus sequences of over 1100 individuals, we delimited eight separately evolving species of *Cichla* rather than the 15 described. Among species we found evidence of rare but widespread introgressive hybridization, while within these species we observed evidence of long-term gene exchange and constrained evolutionary trajectories. In most cases of hybridization, mtDNA was exchanged, while nuclear introgression was extensive to negligible. We estimated a phylogeny among *Cichla* species using sequences of 21 putatively unlinked and single copy nuclear loci. We observed minor discord among loci with respect to the concatenated tree, but when a phylogeny was estimated from separate locus genealogies while accounting for individual uncertainty, a congruent tree was reconstructed. We inferred that less variable loci contributed important phylogenetic information that altered the final phylogenetic estimate. We used extensive mtDNA datasets from three species of *Cichla* that are found in the Orinoco and Amazonas River basins to test the hypothesis that the Casiquiare River permits exchange of genes between populations in each basin. We inferred that either gene exchange has been occurring between these regions or has ceased very recently. Based on the inferred species relationships and using time-calibrated and coalescent-based analyses of unlinked loci, we inferred that the polyphyletic arrangement of mtDNA in *C. orinocensis* was not consistent with deep coalescence and most likely derived from a history of ancient introgression.

Acknowledgements

Considering who and how to properly acknowledge for contribution to my doctoral work makes me think of John Donne's poem 'No man is an island': the list seems to go on and on, and reminds me just now indebted I am. But some individuals stand apart, and I will acknowledge them here in no specific order.

For as long as I have been working on *Cichla*, I have depended on the help and advice of Izeni Farias and Tomas Hrbek. Izeni has given selflessly of her time, and endured many blunders on my part, in order to foster our collaboration. Without her, my work with South American fishes would not encompass one tenth of its present volume, and my perspective on international collaboration would be wholly different. Similarly, Donald Taphorn provided a home and base for me during my excursions in Venezuela, and we had at least one experience we both won't soon forget. In addition to these, several South American graduate researchers have been pivotal to my research. Principally, these were Daniel Toffoli Ribeiro, Valeria Machado, Natasha Meliciano, and Carmen Montaña, not to mention their partners and families. These individuals allowed me to drag them away to the field, away from their own research and obligations, often for weeks or months at a time, to serve largely as a chaperone for a bumbling American researcher. I can only hope that they benefitted from our time together even a fraction of the amount that I have.

Of course, I did not arrive in Brazil or Venezuela without help. Many years ago, David M. Schleser showed me that anyone can discover the mysteries of the Amazon as long as they are willing to get their feet wet. He felicitously pointed me to Kirk Winemiller, at Texas A&M, in whose lab I learned how to be a scientist. Kirk accommodated and supported me in so many ways, and gave me room to discover my own direction in biology. Many of Kirk's students contributed tissues to this study, and I am also indebted to one, David Hoeinghaus, for the generous company he provided in one of my field expeditions. Of Kirk's student though, above all, Hernán López-Fernández deserves credit (or blame) for my pursuit of evolution. Hernán fielded a great many of my questions and musings about evolution and phylogenetics during my undergraduate days, and put up with a headstrong kid who was always sure of the

veracity of his opinion. He has since supported me in various ways throughout my graduate work, and helped me in more than one field trip. I suspect Nathan Lovejoy, on whom I depended (inflicted myself) during my Master's degree and through several joint publications, has similar nightmares about me during those days.

Although I've had the opportunity to visit a great many places in South America in order to collect fishes, it would have been impossible for me to collect all of the localities necessary for a study like the one presented here. I am grateful to the many researchers who contributed tissues to my project. In particular, Paul Reiss provided samples from many important localities that were quite out of the way. I think that he, like Kirk and I, shares a love of peacock bass, and a genuine desire to understand their ecology and evolution.

While my projects required a great deal of collection, they required even more time in the lab with molecular data collection and computer analyses. The students and post-docs of the Ortí lab at UNL and GWU deserve great thanks for teaching me new techniques, discussing the best ways to analyze things, and generally keeping me sane. Jeremy Brozek deserves special credit for the latter, or perhaps more accurately, directing my mind's decent into chaos. Chenhong Li, who preceded me in Guillermo's lab at UNL, has been instrumental in helping me figure out how to collect, analyze, and more basically, understand, large and heterogeneous datasets of the type required by modern systematics. Jason Macrander was a great companion during my last year in Nebraska, commiserating with me about the trials of teaching Biology 101 lab. Jason also contributed to a critical part of my research in spearheading the development and testing of the microsatellite loci. All of my data would have been useless, though, if not for Jean-Jack Reithoven and Sabina Mandahar, who administrated the Bioinformatics computer cluster at UNL. I cannot begin to count the number of times I pestered them with requests to install programs or help me get an analysis working.

During the year I spent at George Washington University, I relied upon a great many people who were welcoming and supportive. Ricardo Betancur has been a great resource for questions about analytical methods and in general for conversations about how to collect and manage molecular data. At the Smithsonian, Rich Vari facilitated my work at the Museum Support Center, which was critical for collection of the

microsatellite data. At the MSC, Jeff, Sarah, Robin, and Maggie gave generously of their time to help me to work efficiently and effectively in a foreign environment.

More basically, I must acknowledge the support I received from my advisors and committee at UNL. They were graciously willing to give me the benefit of the doubt time and again, and offer constructive criticism and advice to improve my research. Diana Pilson went out of her way to facilitate my research during the year I was in Washington D.C., and graciously agreed to act as my advisor after Guillermo moved. Guillermo himself put enormous faith in me, and was always willing to champion my plans and facilitate my goals. In addition to guidance in planning and executing my research, he was generous with financial support to make it possible for me to work in South America, collect the reams of data I wanted, and to make it possible for me to follow him to Washington, D.C. It has been invaluable to have as cosmopolitan and pragmatic an advisor as he.

Finally, it should not go without saying that I would not have found the strength to make it to the end without my family. My wife, Melanie, endured my excited ramblings about phylogenetics with patience, responded to my depression about roadblocks in my research with understanding, and forgave my many broken promises to limit the time I spent on work. My son, Aiden, never complained when I was obligated to work rather than spend time with him, and has been patiently attentive when I relentlessly try to teach him about the natural world. My parents, Don and Lyn, patiently looked away after the numerous time I spilled aquarium water on the floor, tirelessly supported me while I searched for my vocational calling, and eventually found the will to be excited about having their son be a 'fish-ologist.' And my sister, Anne, never made fun of me, at least about preferring animals over most people.

To all of these people I am greatly indebted. This research was also supported by grants from the US National Science Foundation, UNL School of Biological Sciences, UNL Initiative in Ecological and Evolutionary Analysis, UNL Research Cluster, Sigma Xi, Society of Systematic Biologists, and American Society of Ichthyologists and Herpetologists. Work in Brazil was authorized under permit for collection No. 031/2003, 045/IBAMA, 148/2006-DIFAP/IBAMA, permit for access to genetic resources in Brazil No. 034/2005/IBAMA, and Permanent IBAMA License 11325-1/2007.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Our planet, Earth, is at this time unique in our galaxy, and perhaps far more extensively, for its possession of living creatures. Equally amazing is that these creatures, in all their myriad forms, descended with modification from a single ancestral population as a result of errors in genetic rendition. If we are to understand our origins through this process of biological evolution, we must ground our knowledge with the history of this descent, or a phylogenetic tree of life.

Species form the basic building blocks of this phylogenetic endeavor. While populations within species are able to exchange pieces of genetic material, thereby eroding the foundations of phylogeny among them, species develop barriers to interbreeding and thus acquire an enduring set of ancestor-descendant relationships that can be represented by a phylogeny. That being said, the inference of phylogeny, and the identification of species themselves, is by no means an easy task. Modern techniques require a large amount of data and complex analytical methods, both of which involve assumptions about how the data evolved.

In this dissertation, I focus on the inference of species boundaries, phylogeny, molecular evolution, and biogeographic history of a group of freshwater fishes from South America, the genus *Cichla*. South America has the biggest river, the largest forest, the longest mountain chain, and the greatest diversity of organisms of any continent on Earth. It is not entirely clear, however, how all of these phenomena have interacted, and there is no salient hypothesis for how this mega-diversity has accumulated (Hoorn et al. 2010). Any general theory of diversification on Earth must provide an explanation for how the South American tropical diversity has originated. Fundamental to this is an estimate of relationships and ages of the extant organisms in this region.

My analysis of *Cichla* is described in several chapters that are each arranged to be published as separate papers in peer-reviewed journals. Although I wrote these chapters and conducted the analyses, I have listed several authors who have contributed to the study in each case. Modern biology, and science in general, is frequently a collaborative venture, and the multiple-authorship of these manuscripts reflects the logistical or philosophical assistance I received in my pursuits.

The present chapter introduces some concepts not explained at length in the other chapters. Specifically, I describe some of the data sources used in the following chapters, explain the details of analytical methods, and discuss many of the assumptions that these analyses make about the data. Chapter 2 describes the estimation of separately evolving species in this genus by focusing on genetic exchange between subpopulations ascribed to putative species. This is done using several different types of data and analytical methods with different strengths and assumptions. In Chapter 3 I estimate the evolutionary relationships among species and evolutionary significant units of *Cichla* using multiple, putatively unlinked and single-copy nuclear loci. I examine the variation in phylogenetic inference when unlinked loci are assumed to have the same or separate genealogical histories and when the number of loci utilized changes. Chapter 4 examines the biogeographic history of *Cichla* with respect to a unique geographical feature, the Casiquiare connection between the Amazonas and Orinoco River drainages. This paper was published in 2010 in the journal *Molecular Ecology*, and was written well before the other chapters. This temporal disjunction explains minor conflicts between this chapter and the others (e.g. species tree topology, species nomenclature, non-inclusion of microsatellite data). The last chapter, Chapter 5, describes a test of hypotheses for the polyphyletic mitochondrial DNA pattern of one species of *Cichla* (*C. orinocensis*). It distinguishes the capture of the mitochondrial DNA of another species from the retention of ancestrally polymorphic DNA inherited from a progenitor species.

Microsatellites: structure and analysis

A major focus of Chapter 2 is estimating the exchange of genes using a type of marker called microsatellites. Microsatellites are regions of DNA where a small number of nucleotide bases, usually one to six, are repeated in sequence. An example of a di-nucleotide microsatellite would be CACACACA, also denoted $(CA)_4$. Microsatellites can occur in gene exons, although they are more common in non-coding regions such as introns or intergenic DNA, and polymorphisms are therefore generally assumed to be selectively neutral (Li et al. 2002). It is usually expected that microsatellites evolve relatively quickly compared to other DNA loci through a mechanism known as slip-strand mispairing, a kind of error that occurs during DNA duplication. This phenomenon

occurs when the new DNA strand 'slips' or 'jumps' up or down along the repeat region while maintaining nucleotide pairing between some repeats in each strand. The results is that the new DNA copy can exhibit more or fewer repeats than the original strand, in effect increasing or decreasing the size of the repeat region (Ellegren 2004). Microsatellites with large, perfect repeat regions are expected to mutate more quickly than small regions or imperfect motifs because of their greater susceptibility to this error. In addition, other mechanisms such as crossing-over during meiosis may also contribute to microsatellite variation (Li et al. 2002). In any event, their generally rapid mutation rate means that microsatellites often exhibit a high degree of variation within populations, and can be useful for estimating relationships among individuals and gene flow among localities (Goldstein and Schlötterer 1999). In addition, because most of the information in microsatellite loci comes from the size of the fragment (i.e. number of repeats) rather than in the DNA sequence *per se*, microsatellite data have traditionally been more easily and cheaply collected than for loci that require DNA sequencing.

However, the use of microsatellite data is not without its limitations. For estimates of parentage and gene flow, a major limitation is size homoplasy in the data. This occurs when two microsatellite alleles of different size (and history) independently mutate to the same number of repeats. Although some studies have expressed success in using mutations in the DNA regions flanking the microsatellites to identify size-homoplasious alleles (Hey et al. 2004), most times it is impossible to know what proportion of microsatellite variation is effected by this phenomenon (Ellegren 2004). In addition, using mutation models to estimate the relatedness of alleles and probability of size homoplasy beyond the simple infinite alleles model can accommodate some of this variance. Two frequently-utilized examples are the stepwise and two-phase models, which consider genetic distance between alleles based on the fragment sizes and allow for different rates of mutation between size classes, respectively (Goldstein and Pollock 1997). Another limitation of microsatellite data is the observation of biases in their mutation patterns. For example, studies have noted greater mutation rates for larger alleles than smaller, limits to the size a repeat region may take, a greater tendency to increase rather than decrease in repeat length, and variation in mutation rate among loci

(Rubinsztein et al. 1999; Ellegren 2004). Depending on the analytical method, these phenomena can introduce a significant degree of noise or bias into the results.

The microsatellite data in this dissertation were analyzed using two Bayesian clustering programs (STRUCTURE and STRUCTURAMA) that make similar assumptions about the nature of the microsatellite data. These programs both cluster individuals such that microsatellite allele frequencies provide the best fit to a model of Hardy-Weinburg equilibrium (e.g. $p^2 + 2pq + q^2 = 1$ for two alleles, where p and q are the frequencies of the respective alleles). They both implement an infinite alleles model that assumes that each mutation in a locus creates a new, unique allele (Kimura and Crow 1964). This model implies that the exhibition of an allele by an individual reflects the inheritance of that allele, without change, from its ancestor. Thus, the possession of the same allele by two or more individuals implies relatedness between them, and distribution across localities implies genetic exchange or recent co-ancestry. Through strict implementation of the infinite alleles model, these programs provide no accommodation of allele size homoplasy in the dataset. The impact of this bias likely depends on the genetic structure of the populations being analyzed and the degree of variability within and among loci in the analysis. These programs also assume that each locus in the dataset provides an independent estimate of relatedness among individuals (i.e. that alleles among loci are unlinked). A degree of linkage among loci would artificially inflate the homozygosity among individuals and lead to estimates of smaller, more isolated populations by clustering. The typical test for linkage among loci involves the genotyping of a large number of individuals from a presumably panmictic (freely interbreeding) population and testing for deviations from Hardy-Weinburg equilibrium among alleles across loci. Although not included in our manuscript, tests of the loci utilized in this dissertation were made in two divergent species, with both with nonsignificant results (see Master's thesis of J. Macrander).

Inferring haplotypes of nuclear loci

Several chapters utilize the DNA sequences of nuclear loci (apart from microsatellites). Unlike the marker historically favored for phylogeography, mitochondrial DNA, nuclear DNA is diploid, and each individual possesses two alleles

for each locus. However, traditional sequencing approaches do not directly allow for the observation of these alleles separately, and analyses utilizing the DNA sequences of both alleles require additional steps. Apart from limited circumstances where allele size or single-strand electrophoretic mobility allows alleles to be excised separately from a gel, there are two approaches to obtaining these separate sequences, also known as phasing: molecular and computational.

The molecular approach utilizes bacterial sub-cloning or next-generation sequencing techniques. In bacterial sub-cloning, the PCR product from one locus, including both alleles, is mixed with a circular vector that contains an anti-biotic resistance gene, and these are ligated (attached) together. This ligated vector mixture is then introduced to an aliquot of bacteria that is then plated with antibiotic media. Due to the asexual and clonal growth of these bacteria, those containing the vector grow into colonies containing only a single allele. The colonies are then picked and sequenced separately to obtain each allelic sequence. However, there are two major limitations to this approach. One is the time and expense of cloning, and the difficulty of obtaining colonies with each allele, which can be considerable. The second and more important limitation for this technique is the errors introduced in bacterial replication. Mutations are introduced in the cloned sequences through reduced proof-reading efficiency of bacterial polymerases, approximately one to three per thousand base pairs (personal observation), and at the levels of sequence divergence evident among closely related species, this error can mimic the true differences among alleles. One way to accommodate this error is to sequence multiple clones for each allele, expecting that only those mutations shared by all clones will reflect true variation among alleles. This was the approach we utilized when necessary, although most nuclear phasing was accomplished via the computation approach. The other molecular strategy, next-generation sequencing approaches, utilize a step in which PCR proceeds from a single DNA copy. In this way, the DNA sequences produced correspond to a single allele. However, the practical utility of this approach is currently limited by the ability to identify the correspondence of sequences with specific individuals, and the read length (sequence length) produced by the machinery. Also, any PCR recombination among alleles (PCR hybrid artifacts) would be perpetuated in the final data.

An alternative approach to the molecular strategy is to use the genotypic data among individuals produced by direct sequencing techniques to estimate the phase of the alleles. In this dissertation, this computational phasing was accomplished using the Bayesian program PHASE. Essentially, this program models the distribution of mutations in an extended coalescent among alleles (see below). The allelic phases can be estimated with or without recombination, and known phases (such as from a limited number of cloned sequences) can be incorporated to inform the analysis. The MCMC chain estimates probability of haplotype assignment and phase for each mutation, and a cut-off value for phases below a certain probability can be specified. The program is most effective when there is a low degree of independent mutations to the same state among sequences, as is generally the case with low-variability data, and when many individuals are utilized in the analysis (personal observation). However, as with any computational approach, there is a certain error rate for the phasing results. A recent review suggested that this rate is relatively low, but is more frequent for rare and more divergent alleles (Garrick et al. 2010). The effects of incorrect phasing are likely to be larger where the analysis depends more on the topology of individual genes, and where models (e.g. coalescent) depend on the accurate estimation of branch lengths or divergence times.

Phylogenetic inference

Each of the following chapters makes a phylogenetic or genealogical estimate using probabilistic methods with one or more loci. The partitioned or separate locus phylogenetic analyses used here assume several other important things: 1) that sites in an alignment (locus) evolve independently and in identically distributed patterns (i.i.d.); 2) that evolution has been tree-like, and can accurately be represented as a bifurcating tree; 3) that mutational processes (patterns across sites, rates at individual sites) are stationary across the tree, and do not change between branches or from root to tips; 4) that mutations are reversible, that is, that a mutation to one state does not bar the mutation to any other state; and 5) that mutation occurs by a Markov process, wherein future mutation only depends on current state, not previous states for that site (Whelan et al. 2001; Felsenstein 2004; Kelchner and Thomas 2007). Each of these assumptions is likely violated to some degree in most phylogenetic analyses. For example, in Chapter 3 there is

a discussion of appropriate partitioning schemes for large datasets. This partitioning is meant to help the analysis to conform to the assumptions of i.i.d. by allowing sites with different patterns to be analyzed under multiple models. But as discussed in the chapter, it is impossible to know which sites truly evolve under the same processes, and our stochastic and *a priori* process partitions are likely to only accommodate the largest variations. Similarly, it is unclear that DNA polymorphisms always conform to tree-like evolution. Particularly concerning intraspecific polymorphism, recombination and extant ancestral alleles can violate these assumptions and lead to statistical inconsistency in phylogenetic estimates. Any violations in these assumptions will have some affect on the inferred molecular model (topology, branch lengths, substitution frequency matrix, nucleotide frequencies, mutation rate). Depending on how strong the phylogenetic signal from the data is otherwise, some degree of violation, such as a small amount of recombination or substitution frequency variation, can generally be accommodated by parametric models. Further, our analyses were largely Bayesian, which integrate over uncertainty in estimates of one or more parameters. Where these violations tend to have the greatest effect is at very short branches, or where accurate branch lengths are crucial (Kelchner and Thomas 2007).

Another crucial assumption of the phylogenetic analyses in this dissertation was that the sites in each matrix were homologous. Homology describes characters that are inherited from a common ancestor, and whose similarity reflects that descent. Specifically, this means that sequences or alleles at each locus are assumed to derive from orthologous gene copies. Orthologs are those genes of which there is a single copy in the genome, and are duplicated only by organismal lineage divergence (speciation) events. Paralogous genes, on the other hand, are those gene copies of which there is more than one homolog within a single genome, and have their origin in genome duplication or translocation events (Koonin 2005). It is generally not an easy task to detect paralogous gene copies, but two factors worked in our favor for the current dataset. First, we analyzed data from a closely related group of species that reflected a time period over which gene duplication was not likely to have been frequent. Second, we avoided any genes that exhibited the properties of paralogous genes, such as sites that were heterozygous in all sequenced individuals (potentially indicating the polymorphism

among paralogs). Another assumption of homology regards the alignment of nucleotide sites among sequences. For loci that did not vary in length among individuals, this is a trivial matter. However, for length variable alleles, this matter has received great attention in the literature. In our analyses, we used the latest methods of sequence alignment and avoided the use of those DNA regions that were alignment ambiguous.

Within this framework of phylogenetic or phylogeographic analyses, we utilize several statistical tools in several chapters. One of these is the likelihood ratio test (Huelsenbeck and Rannala 1997). This technique compares the likelihoods of nested models (models in which the comparison is between variants where one or more parameters is fixed versus free) against the differences expected based on a X^2 distribution. We utilize this, for instance, in comparing phylogenetic models with and without the assumption of ultrametricity (the molecular clock) and in comparing the coalescent models of divergence between populations with and without gene flow. A similar technique is our use of the Shimodaira-Hasegawa and Kishino-Hasegawa topology tests. These techniques compare the likelihoods of alternative phylogenetic models (topologies). Unlike the LRT, however, these analyses create their own test distribution of likelihoods. The difference in model likelihoods is then compared to this distribution, as above. To choose mutation models for individual loci, we utilized the Akaike information criterion (AIC; Akaike 1974). This statistic rewards models that show a higher likelihood given the data, but also penalize those models with more parameters. In this way, the AIC helps to avoid overparameterization by striking a compromise between likelihood and parameterization. We similarly used the Bayes Factor comparison to choose between competing models. The Bayes Factor, however, is used in a Bayesian framework, where uncertainty in model parameters is accommodated by the MCMC framework. This technique does not have an explicit penalty for additional parameters; rather this is intergrated into the Bayesian analysis. Each parameter in a Bayesian analysis is given a prior distribution, which for any value is a number less than one. Additional parameters, therefore, decrease the prior likelihood even while the posterior probability may increase. Therefore, to be favored by the Bayes factor, a model must exhibit a significant increase in data fit to overcome the decrease in prior likelihood (Raftery 1996; Nylander et al. 2004). However, the Bayes factor is not a formal

hypothesis test with a test distribution and p-value. Model choice using the Bayes factor generally follows the guidelines of Kass and Raftery (1995).

Coalescent models

Analyses in several chapters used a population genetic model known as the coalescent (Kingman 1982). This model describes the descent of orthologous gene copies (alleles) from a single common ancestral gene copy through time. It predicts, based on the effective population size of the population, what time in the past two gene copies are likely to have shared a common ancestor. When population size is unknown, the distribution of mutations among gene copies from individuals in a population can be used to estimate population size and time to coalescence (Rosenberg and Nordborg 2002). Several of the models used here are extensions of this model where the coalescence in one or more extant and ancestral populations is modeled (e.g. isolation with migration). For example, if two genetic polymorphisms in an extant population (or species) fail to coalesce before the time at which that population diverged from its most closely related (sister) population, the coalescence of those gene copies can be modeled in the population ancestral to both of those extant populations by estimating an effective population size for the ancestral population. Although relatively simple in theory, the estimation of these parameters is quite difficult computationally and often results in wide confidence intervals for parameters of interest. We worked around this phenomenon by making multiple runs of the coalescent models, using somewhat more informative priors in our mutation models and placing bounds on prior distributions based on preliminary runs. We also used these models in ways that were not dramatically affected by moderate confidence intervals for some parameters, such as hypothesis tests and model comparisons.

One important assumption of these coalescent models, as well as the phylogenetic analyses using separate loci (i.e. Bayesian concordance analysis), is that each locus included in the analysis is unlinked via recombination during meiosis. This means that each gene has a separate genealogy, and each genealogy provides an independent estimate of coalescence time, population size, and or phylogeny. However, with non-model organisms for which a genome sequence or genetic map is not available, linkage

between gene regions cannot be estimated easily. Fortunately, the odds of linkage between a random set of loci decrease as the number or size of chromosomes in the genome increases. Since our 21 loci are derived from a genome of 48 non-degenerate chromosomes, the odds of two being found separated by a short distance on one of these chromosomes is small. If, however, there were a degree of linkage between two or more loci, this would artificially decrease the population sizes estimated from these loci, and bias the estimate of phylogeny towards the genealogy exhibited by these loci. The phylogenetic (multispecies coalescent) models further assume that the only process governing coalescence among species is genetic drift, not gene flow (Carstens and Dewey 2010). If introgression had facilitated gene exchange between these species after divergence, it would artifactually increase the population sizes and decrease the time since divergence estimated between these species. This would be particularly troublesome where the hybridizing species were sister species, as there would be no topological model with which to distinguish deviations (see Chapter 5). While it is impossible to know if two species have hybridized in the past without leaving an identifiable signature, we avoided the use of individuals from localities where hybridization was inferred in our analysis of species phylogeny.

Summary

"All models are wrong, but some are useful" (G.E.P Box, 1979). By identifying the assumptions of each model and analysis, and understanding how violations of those assumptions may affect the results, we can extend the utility of those methods. This dissertation profits by applying multiple datasets, models and analyses to investigate evolutionary processes while accommodating uncertainty and minor violation of assumptions. Fundamentally, we will never be able to fully deduce the history of life, but we can at least be explicit about what we do not know.

Chapter 2: Species delimitation and examination of hybridization in Amazonian peacock cichlids (genus *Cichla*) using mitochondrial and nuclear sequences, and microsatellite genotypes

Authors: Stuart C. Willis, Jason Macrander, Izeni P. Farias, & Guillermo Ortí

Introduction

Species hold a pivotal role in biology and our understanding of evolution. Contributors to the New Synthesis emphasized the distinction between species, which are not only the product of evolution but also function directly in the evolutionary process, and those taxa above and below the species rank (Simpson 1961; Mayr 1963). Species are generally considered to be groups of interbreeding individuals (populations), which exchange genetic material with minimal functional constraint and more exclusively with con-specifics than with other groups of individuals, and as a result, show more phenotypic (morphological and functional) similarity and experience a constrained group-wise evolutionary trajectory (sensu reviews by de Queiroz 1998; Coyne and Orr 2004). As a result, individuals ascribed to a given species are often treated interchangeably in an array of biological investigations (Hull 1977; Apagow et al. 2004; Isaac et al. 2004; Padial and de la Riva 2006). However, two problems have plagued the delimitation of species units. First, there is a continuing debate in the literature as to what species are, a debate which seems to center on the emphasis of operational criteria for defining species versus the functional roles of species in evolutionary or ecological processes (Mishler and Donoghue 1982; Coyne et al. 1988; Cracraft 1989; Templeton 1989; de Queiroz 1998; Hey et al. 2003). A second, more pragmatic issue is simply the difficulty of discerning species boundaries in nature, a challenge that follows not only from the former conundrum, but also from the inadequacy of current tools to provide a sufficient understanding of natural processes (Wiens and Penkrot 2002; Morando et al. 2003; Sites and Marshall 2003; Pons et al. 2006; Knowles and Carstens 2007; O'Meara 2010).

One process that has been emphasized repeatedly in the literature across many species concepts is genetic exchange or gene flow (Dobzhansky 1937; Mayr 1942; see also de Queiroz 1998; Coyne and Orr 2004). Genetic exchange among individuals across subpopulations is important because it connects groups of individuals demographically, constrains the rate at which adaptive variation may spread, and creates an extended and interconnected set of ancestor-descendant relationships across time (a lineage) (de Queiroz 1998; Sites and Marshall 2003). We explored this paradigm of gene flow between subpopulations to delimit species in a widespread group of Neotropical fishes (we consider a subpopulation to be a group of individuals in an area defined by a species-specific scale that implies probable panmixia; operationally these are our sampling localities). We used multiple unlinked genetic markers to estimate the qualitative pattern of current and historical gene flow between groups of individuals from many different localities. We did not, however, apply a strict cutoff for levels of gene flow necessary to lump or split subpopulations together, using the totality of results to make estimates of species boundaries. The problem with a strict criterion is that there will always be species groups in which it fails (Hickerson et al. 2006). A good example is the criterion of species monophyly for loci such as the mtDNA (Avise and Ball 1990; Baum and Shaw 1995), which has been criticized as being overly restrictive and overly permissive in turn (Moritz 1994). Rather, it seems more holistic and effective to compare the patterns of historical and contemporary gene flow represented by multiple markers and look for the congruence among patterns, whatever degree it may be, that denote separately-evolving lineages.

Patterns of gene flow are a logical criterion for understanding species limits, but this is not without its limitations. For instance, genetic disjunctions between localities resulting from a high degree of spatial genetic structure within a species (e.g. isolation by distance) may be misinterpreted as separate meta-populations (i.e. species). It is therefore important to sample densely enough on the organism-specific scale to observe the connectivity between subpopulations (Morando et al. 2003). Further, it will often be impossible to determine an adequate sampling strategy entirely *a priori*. Rather, it will frequently be necessary to sample repeatedly with a focus on testing observed genetic discontinuities. Similarly, while gene flow connects subpopulations together into a meta-

population, it may also connect meta-populations together (introgressive hybridization). As a result, depending on the frequency of hybridization, it may be a challenge with this criterion to distinguish gene flow within species from introgression between species. We tentatively expected hybridization, when present, to be represented as a much smaller proportion of gene exchange between groups of subpopulations than genetic overlap within those groups (Shaw 2001). Based on our putative identification of hybridization, we treated this investigation as an opportunity to examine the frequency, circumstances, and genetic impact of hybridization across species.

Our analysis focused on Neotropical freshwater fishes in the genus *Cichla* Schneider, 1801, commonly known as peacock basses. These cichlid fishes are large bodied (up to 12 Kg), diurnal piscivores that play critical ecological roles in fluvial ecosystems of tropical South America (Winemiller and Jepsen 1998; Winemiller 2001; Layman and Winemiller 2004). Given their large size and voracious strike, *Cichla* are highly exploited as food and sport-fishing resources throughout their range. Tagging studies in their native distribution have shown that most individuals are highly site-fidelous, even across years, but do exhibit infrequent long-distance dispersal (Hoeinghaus et al. 2003). They are seasonally monogamous and show extensive parental care including mouth brooding (Winemiller 2001; P. Reiss, pers. obs.). There are currently 15 recognized (putative) species of *Cichla*, all of equal karyotype (2N=48). This genus includes both sympatric and allopatric taxa (Kullander and Ferreira 2006), although allotopy and resource partitioning are common (Jepsen et al. 1997). Because body shape has a high degree of conservation in this genus, due presumably to the functional constraints of nearly-complete piscivory as adults (Collar et al. 2009), species discrimination is often based on subtle color pattern differences, mean differences in meristic variation, and geography. Further, previous work in *Cichla* using mitochondrial DNA showed both congruence and incongruence with morphological estimates of species (Willis et al. 2007). In addition, studies examining morphological and mitochondrial mismatch, as well as karyology, have inferred hybridization in natural fluvial (Willis et al. 2007) and artificial reservoir (Brinn et al. 2004; Oliveira et al. 2006) environments.

The Neotropics are home to the most species-dense assemblage of vertebrates on earth (~10%), and the largest assemblage of freshwater fishes (Reis et al. 2003). While

there are many genera of Neotropical fishes with many more species than *Cichla*, and in much greater need of species delimitation, we chose to examine this genus because they are well-known, they are wide-spread, and there was likely to be a reasonable number of species for analyses with a high data requirement. Our null hypothesis was that molecules would delimit the same species recognized by morphological review (Kullander and Ferreira 2006). We emphasize that our conclusions about species limits and identifications of introgression must be considered working hypotheses. Nevertheless our dataset and analyses provide an in depth and insightful look into the nature of widespread species in a mega-diverse region. Our results from *Cichla* provide a set of null-expectations for species diversity in other Neotropical fish groups.

Methods

Sampling and DNA extraction

We collected fin, gill, or white muscle tissue from fishes in the Amazonas, Orinoco, Essequibo, and Maroni River drainages and preserved it in 90% ethanol or DMSO-EDTA saturated with NaCl (Table 1, Figure 1). Permit numbers are listed in the Acknowledgements. We identified each individual according to Kullander & Ferreira (2006) using morphology, as possible. While many vouchers were taken (information available upon request), where possible fishes were photographed, sampled non-destructively (dorsal fin), and released alive. Sampling was done between 2003 and 2010 with new locations chosen to cover most biogeographically important areas (those potentially harboring additional species) or to test observed genetic discontinuities when possible. We targeted to collect at least 10 individuals per locality where possible. DNA was extracted from samples using the Qiagen DNeasy extraction kit (Qiagen Inc.) following manufacturer's recommendations.

Molecular Markers

We collected data from three different sources. First, for every sample we sequenced the mitochondrial control region (mtCR), and for many samples, we also sequenced the mitochondrial ATPase 8,6 gene (mtATP). The mtCR, with among the fastest mutation rate outside of tandem repeat regions, provides an unambiguous

assessment of genealogical connection between individuals in order to estimate gene flow and introgression (Morando et al. 2003). In addition, the faster rate of lineage sorting relative to nuclear loci means that mtDNA has a higher chance of representing recent gene flow rather than incomplete sorting of ancestral variation (Avise et al. 1987). However, some properties of the mtDNA may cause it to give a biased signal of gene flow (e.g. selective sweeps, etc.) (Gerber et al. 2001; Galtier et al. 2009). Moreover, there have been an increasing number of recent reports that mtDNA often crosses species boundaries more readily than nuclear loci (Chan and Levin 2005; Linnen and Farrell 2007; but see Di Candia and Routman 2007). As such, although it can indicate the frequency of hybridization and the opportunity for transfer of adaptive variation (Anderson and Stebbins 1954; Lewontin and Birch 1966), it may not accurately represent overall genomic introgression. Moreover, any single marker may provide a biased or misleading signal of genetic exchange (Funk and Omland 2003). In order to further test patterns of gene flow, we also obtained nuclear data of two types. First, we sequenced two nuclear loci consisting of both open reading frame and intron segments: a tyrosine kinase gene (Xsrc; Sides and Lydeard 1999) and the micropthalamia b receptor protein (Mitf; Won et al. 2006). We sequenced these loci from a subset of all individuals. Because these loci have slow mutation rates, they are generally expected to show a low degree of homoplasy (Zhang and Hewitt 2003). However, the longer coalescence time of nuclear loci means that it can be difficult to distinguish gene flow and hybridization from the sorting of ancestral polymorphism among recently diverged species (Nei and Li 1979; Pamilo and Nei 1988; Maddison 1997). Second, we genotyped most individuals for a panel of 12 microsatellite loci. This source of nuclear data hypothetically suffers the same problem as the sequenced nuclear loci, a slow rate of coalescence. However, while lineage sorting of any nuclear locus may be a slow process, deviations from Hardy-Weinberg and linkage equilibrium at these hypervariable loci can occur between species over many fewer generations (Goldstein and Schlötterer 1999). On the other hand, these loci are suspected to exhibit biases in their mutation patterns that create an unknown degree of homoplasy, potentially reducing their usefulness for estimating gene flow (Rubinsztein et al. 1995; Rubinsztein et al. 1999; Ellegren 2004; Hey et al. 2004). Thus, while each of these datasets has weaknesses, we used them in combination to make a

more accurate estimate of subpopulation connection and species boundaries, expecting that their collective strengths would help counter their individual weaknesses.

We collected data from the mtCR (~550bp) and mtATP (842 bp) using previously described primers and conditions (Willis et al. 2007; Willis et al. 2010). Many of these data were published previously on Genbank (CR: DQ841819-DQ841946, GU295709–GU295740; ATP: GU295741-GU295801), and new data will be published there as well. In addition, we obtained the sequence data generated by another study of *Cichla* (Renno et al. 2006) in Bolivia and Peru (DQ778661-DQ778712). Sequences were checked and assembled using CODONCODE ALIGNER (CodonCode Corp.).

PCR reactions for Xsrc and Mitf (each ~750 bp) contained 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2.0 mM $MgCl_2$, 200 µM each dNTP, 0.1 µM each primer, 1.5 µL of 20 mg/mL bovine serum albumin (Fermentas), 0.5 U of Takara ExTaq polymerase (proof-reading exonuclease activity), and 3 to 4 µL DNA extract (10–50 ng⁄µL) in 30 µL reaction volumes. Thermal cycling conditions for Mitf on an MJ Research PTC200 thermal cycler were 1 min at 94°C, 35 cycles of 30 sec at 94°C, 30 sec at 54°C, and 1.5 min at 72°C, followed by 10 min at 72°C. Thermal cycling of Xsrc required a touchdown protocol of 1 min at 95°C, 30 cycles of 15 sec at 98°C, 30 sec at *X*°C, 1.5 min at 72°C, followed by 10 min at 72°C, where *X* was 64°C for 3 cycles, 62°C for 3 cycles, 60°C for 3 cycles, 58°C for 6 cycles and 52°C for 15 cycles. PCR products were sequenced at the University of Washington High Throughput Facility. Chromatograms were checked and assembled using CODONCODE ALIGNER. Most sequences were estimated using direct sequencing, except in cases where individuals were heterozygous for an indel on each allele (or otherwise difficult sequences), where we used bacterial sub-cloning and sequenced 5-10 clones to estimate the correct genotypes. We estimated haplotype phase and identity among individuals using the recombination model of PHASE (Stephens et al. 2001) and a phase probability of 0.6.

Each of our 12 microsatellite loci had a di-nucleotide repeat motif. Tests of linkage between our microsatellite loci have been examined previously (Macrander, Willis et al. in preparation); primer sequences and thermal cycle conditions are available there as well. We made no attempt to test HWE or LD in individual subpopulations here because small sample size would likely create many spuriously significant deviations.

PCR reactions for the microsatellite loci included 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 1.5 mM $MgCl_2$, 200 µM each dNTP, 0.25 µM each primer, 0.24 µL of 10 mg/mL bovine serum albumin (New England Biolabs), 0.5 U of taq polymerase (Biolase), and 1 µL DNA extract (10–50 ng⁄µL) in 6 µL reactions. Reactions were assembled in 384-well plates using the Matrix PlateMate 2x2 (Thermo Scientific) and amplified on a MJ Tetrad thermal cycler (MJ Research). Each 384-well plate had at least four positive and four negative control samples. One primer for each locus had one of four fluorescent dyes, and fragment sizes were determined in three runs per sample on an ABI 3730 Automated Sequencer (Applied Biosystems). Genotypes were scored using GENEMAPPER (Applied Biosystems).

Sequence Analysis

MtCR sequences were divided into 4 groups for alignment with the L-INS-i algorithm of MAFFT (Katoh et al. 2005): 1) *Cichla orinocensis*, 2) *C. intermedia*, 3) *C. ocellaris*+*C. monoculus*+*C. pleiozona*+*C. kelberi*+*C. nigromaculata*, and 4) *C. temensis et al*. Importantly, putative species in each of these groups are allopatric, while sympatry is only observed between the groups. We aligned these separately because although the L-INS-i algorithm is highly accurate, it has limitations for the number of sequences that can be aligned at one time. Using these four alignments separately, we identified unique sequences (haplotypes) using TCS 1.21 (Clement et al. 2000), treating gaps as a fifth state. Alignments of the unique sequences were combined together and aligned again using L-INS-i. We did not collapse haplotypes a second time to identify haplotype sharing between groups; rather, these are evident as zero-length branches on our inferred phylogram. Where available, the mtATP sequence of one individual bearing a haplotype was concatenated on to the mtCR sequence. The mtATP sequences did not vary in length and did not require additional alignment. We estimated appropriate models of sequence evolution for each of these two partitions (mtCR, mtATP) in TREEFINDER v. Jan2008 (Jobb et al. 2004). These were HKY+Γ and TN+Γ respectively. We then inferred a maximum likelihood phylogeny under these models in TREEFINDER using 1000 search replicates. We estimated support for major branches using 100 bootstrap replicates in TREEFINDER, but included only those sequences with both mt-genes in this analysis. For

the nuclear sequence loci, models of evolution and maximum likelihood genealogies were estimated for each gene using TREEFINDER as above (Mitf: HKY; Xsrc: HKY+Γ). We did not estimate branch support for these loci due to the restricted number of mutations. Trees were mid-point rooted. Within species, or between species in the case of hybrids, gene flow between subpopulations was indicated when localities or species designations are admixed across branches in the genealogy. Using these three genealogies, we looked for exclusive haplotypes or haplotype lineages that corresponded to putative species. We identified hybrids by mismatch between morphology (species ID) and genetic lineage, clear incongruence between datasets, and/or when particular genetic lineages were found outside of otherwise exclusive spatial distributions.

Microsatellite analyses

To identify gene exchange within species as implied by genetic overlap, we analyzed the microsatellite genotypes using the Bayesian clustering programs STRUCTURE (Pritchard et al. 2000) and STRUCTURAMA (Huelsenbeck and Andolfatto 2007). Both of these programs attempt to match individuals to clusters that best correspond to a model of Hardy-Weinberg equilibrium, a model which implies a high degree of gene flow within clusters but a lower degree of gene flow between clusters. However, where admixture between clusters is indicated, this should also represent gene flow between subpopulations in each cluster. The program STRUCTURE has been extensively applied in tests of population structure (e.g. Lee and Edwards 2008), as well as species boundaries (e.g. Edwards et al. 2008). We divided the microsatellite data into clades B and A (groups 1-3 versus 4 from above, respectively) to avoid the effects of fragment size homoplasy over larger phylogenetic distances. With STRUCTURE, we made 20 runs of the program with K (number of clusters) from 1 to 10. As most of these analytical constructions resulted in an asymptotic increase in the natural log probability of the data given the number of clusters (LnP(D|K)), we used the second order rate of change between runs of different K (ΔK; Evanno et al. 2005) to estimate the optimal value of K. We ran the program with the *r* (locality) prior implemented (Hubisz et al. 2009). This prior ranges from 0 and above, with values between 0 and 1 indicating that locality information is informative for clustering, while values above 1 indicate that it is

not. We made these runs with an initial value of *r* at 1 and a maximum limit of 100. We found that it took much longer for this parameter to converge than is typical for other parameters in STRUCTURE, so each run was made with 1 million sample generations after 1 million generations of burn-in. Evanno et al. (2005) found that their metric, ΔK, identified the optimal clusters at the highest hierarchical level in the data; inferring subsequent structure required dividing the original dataset. Thus after each series of runs, where ΔK indicated discrete clusters (clusters with a low degree of overlap or few individuals with split assignment posterior probability) we divided the data according to the clusters and made another series of runs as above. As this metric cannot indicate K=1 as optimal, we continued to divide and reanalyze the data until the inferred optimal clusters showed a significant degree of admixture, or LnP(D|K) showed a maxima for K less than K=10. For each optimal K for each division, we averaged the posterior probability of individual assignment across all 20 runs using CLUMPP (Jakobsson and Rosenberg 2007).

STRUCTURAMA differs from STRUCTURE in that rather than requiring the user to specify *a priori* the number of clusters to which individuals should be assigned, STRUCTURAMA uses a Dirichlet process prior for cluster assignment, allowing the number of clusters to be a random variable and sampled by the chain as well. With STRUCTURAMA, we ran the two datasets (clade A and clade B) for 100,000 generations with 20 chains and a heating value of 0.03, sampling the cold chain every 100 generations and discarding the first 500 samples as burn-in. In initial tests, we found that runs of 1 million generations produced similar results. We ran each species group's dataset treating number of clusters as a random variable, and 4 runs each for priors of 5, 10, and 15 expected clusters. We did not divide the data for subsequent runs as with STRUCTURE. We found it computationally intractable to run analyses with both admixture in individual assignment to cluster and number of clusters as a random variable. However, the STRUCTURAMA documentation suggests that when the admixture model is used with a fixed number of clusters, the model becomes equivalent to that implemented in STRUCTURE (Huelsenbeck and Andolfatto 2007). We took the median number of clusters across all runs as optimal, and for those runs with this number in the mean partition, we summarized assignment of individuals to cluster across runs using

CLUMPP. To compare admixture for this number of clusters, we examined results from the STRUCTURE runs for this number of clusters (with no prior on assignment of individuals to clusters) summarized using CLUMPP.

In cases of putative recent hybridization, we wanted to examine the degree of introgression of the nuclear genome. We again analyzed our microsatellite data using STRUCTURE, but in this analysis we used only data from the putative hybridizing localities, and nearby non-hybridizing localities. Further, for individuals from non-hybridizing localities, we specified their population of origin and updated allele frequencies in the analysis only using these individuals. The analysis then estimated what proportion of the genome of the putative hybrids was derived from each parental species. This analysis was run for 100,000 generations after 100,000 generations of burn-in, without the *r* prior. We made several runs to confirm proportions across runs, but present the results of a single run. We tested whether or not cluster probability of hybrid individuals was significantly less than a mean expectation of 0.9 or 0.95 using one-sample t-tests.

Results

MtDNA Genealogy

We sequenced the mtCR (aligned 563 bp) for 1130 individuals of *Cichla*, including data from our previous publications. To this we added the 19 haplotypes from the 47 individuals of *C. monoculus* and *C. pleiozona* surveyed by Renno et al. (2006). Removal of redundant sequences resulted in 11 haplotypes in *C. intermedia*, 61 haplotypes in *C. orinocensis*, 154 haplotypes in *C. monoculus* and the remaining clade B species, and 98 haplotypes from *C. temensis* and the remaining clade A species. Overall, these haplotypes ranged from one mutation or gap to 16% uncorrected sequence divergence. Our search for the maximum likelihood genealogy, in which 121 individuals had mtATP concatenated (842bp), resulted in a tree with LnL -9497.713 (Figure 2). The mid-point root and topology of this tree agreed with our previous analyses using a somewhat different mtDNA dataset and partitioning scheme (Willis et al. 2010). In this genealogy of haplotypes, most putative species had exclusive or nearly exclusive haplotypes and haplotype lineages: *C. intermedia, C. orinocensis, C. kelberi, C. ocellaris,*

C. temensis, C. mirianae, C. melaniae, C. piquiti, and *C. pinima*. These species had haplotypes that were a minimum of six (uncorrected) mutations different from any other species, and usually many more. Several putative species exhibited haplotypes that were shared with, or one mutation different from, haplotypes of another species, but these haplotypes made up the minority (<5%) of the total individuals of that species (identified in Figure 2). These were: *C. nigromaculata* (Mavaca, 10 of 10 individuals) that exhibited *C. temensis* haplotypes; *C. intermedia* (Parguaza, 2 of 2) that exhibited *C. orinocensis* haplotypes; *C. pinima* (Tocantins, 3 of 4) that exhibited *C. piquiti* haplotypes; *C. orinocensis* (Preta da Eva, 6 of 6; Novo Airão, 3 of 11) that exhibited *C. monoculus* haplotypes, and *C. monoculus* (Novo Airão, 1 of 11) that exhibited *C. orinocensis* haplotypes. Importantly, in all of these cases these individuals were sympatric with individuals of the species with whose haplotypes they were grouped, or adjacent to localities where they were found. Based on the rapid mutation rate at this locus, the geographic distribution of genetic overlap, and the otherwise high degree of exclusivity of haplotypes in these putative species, we inferred this haplotype sharing to be evidence of recent introgressive hybridization.

Other putative species shared many or all of their haplotypes with other species, or had exclusive haplotypes that were one mutation away from, and nested among, the haplotypes of another species. These included: *C. nigromaculata* haplotypes that nested among haplotypes from *C. monoculus*; and haplotypes from *C. jariina*, *C. thyrorus*, and *C. vazzoleri* that nested among haplotypes of *C. pinima*. This pattern suggests that these species are so closely related as to be indistinguishable. Another pattern was exhibited by *C. pleiozona* and *C. monoculus*. While a large portion of individuals classified as *C. monoculus* (240 of 324, 75%) exhibited haplotypes from a large monophyletic group, all but one fish from the middle Tapajós River (Jacareacanga), middle and lower Madeira River (Aripuanã, Humaita, Cunia, and Canumã), and middle and upper Purus River (Boca do Acre, Labrea, and Tapauá) exhibited haplotypes that were more closely-related to the haplotypes from nominal *C. pleiozona*. The remaining individual from the Canumã exhibited a haplotype from the main *C. monoculus* group. This suggests that either *C. monoculus* and *C. pleiozona* exhibit repeated instances of introgressive hybridization, or that they are not different species.

Finally, some putative species exhibited haplotype lineages that were exclusive to those species (except where described above), but polyphyletic. These included: *C. orinocensis*, which exhibited two mtDNA clades, one sister to *C. intermedia* and the other sister to the *C. monoculus* 'main group'; *C. ocellaris*, which exhibited two clades of mtDNA, corresponding to the Essequibo (plus Takutu) and Maroni River drainages; *C. pinima*, which exhibited 3 clades: one sister to *C. piquiti*, another sister to *C. melaniae* + *C. mirianae*, and another nested among *C. melaniae* + *C. mirianae*; and *C. melaniae*, with one clade very closely related to *C. mirianae*, and a second clade sister to the former clade of *C. melaniae* + *C. mirianae* + the latter clade of *C. pinima* (see Figure 2). In the case of *C. orinocensis*, these two clades were largely allopatric, with one clade found in the Orinoco and upper Negro, and the other clade found in the middle and lower Negro. However, both clades were found together in the geographically intermediate Daraá locality. Similarly, the two main clades of *C. pinima sensu lato* were found together at two localities (Orixinimá and Tapajós mouth), while the third clade was restricted to a single locality (Vittoria do Xingu). Likewise, the two clades of *C. melaniae* were found sympatrically at two different localities (Altamira and Iriri) in the middle Xingu River drainage. These observations suggest that these clades result either from incomplete sorting of ancestral polymorphism (i.e. deep coalescence) or ancient introgression events (Willis et al. 2007). In the case of *C. ocellaris*, the geographic isolation of these lineages implies the presence of multiple evolutionarily significant units in watersheds of the Guyanas region.

Nuclear gene trees

We sequenced 150 individuals for the Mitf gene, which was variable at 24 sites along the 743 bp and resulted in 19 haplotypes. Similarly, we sequenced 139 individuals for the Xsrc gene, which was variable at 32 of the 747 bp and exhibited 36 haplotypes. Maximum likelihood genealogies of these loci were found with an LnL of -1196.07 and -1337.985 for Mitf and Xsrc respectively (Figure 3). In general, we observed few exclusive alleles or allele lineages among putative species, although alleles in the phylogeny did seem to mimic the major phylogenetic structure (Clade A, B1, B2). With so few mutations, it is difficult to distinguish the sharing of alleles within each clade as

either gene flow or the incomplete sorting of ancestral polymorphism. However, these trees are useful for identifying introgressive hybridization between species in clades A and B. In both trees we observed that *C. ocellaris* from the Cuyuni exhibited haplotypes more characteristic of *C. temensis*, while *C. temensis* from the Guri Reservoir on the Caroni River exhibited haplotypes characteristic of *C. orinocensis* (Figure 3). As above, these putative hybrids were either sympatric with individuals of the potential donor species, or adjacent to localities where they were found.

Microsatellite clustering

We genotyped a total of 1034 individuals for the 12 microsatellite loci. Individuals in this dataset had missing data at a maximum of 4 loci, resulting in 0.67% missing data in the overall dataset. In most cases, missing data corresponded to samples with partially degraded DNA. However, the exception was individuals from several localities for *C. pinima*, *C. vazzoleri*, and *C. jariina* that could not be amplified or scored consistently for locus CoriB6.2 despite repeated attempts. This may be indicative of null alleles at this locus, or alleles that do not amplify due to mutations in the priming site. As the presence of null alleles in a heterozygous condition with amplifying alleles can bias estimates of heterozygosity and Hardy-Weinberg equilibrium, we repeated our analyses of clade A without this locus and observed qualitatively similar results. Otherwise, there was a significant variability at each locus in each clade of *Cichla* (Table 2), meaning these loci should be useful for estimating population connectivity. Several of the loci exhibited one base pair differences in fragment sizes, rather than the multiple of two base pair differences expected from di-nucleotide repeats. As these sizes persisted in samples that were re-amplified and genotyped two or more times, we scored alleles according to their electrophoresis mobility and made no attempt to force conformation to a two base pair sequence. We did not include samples from artificial reservoir habitats (nominal *C. temensis* from Guri) in the following analyses.

For the clade A species, *C. temensis* et al., we observed that the probability of the data given K in STRUCTURE (LnP(D|K)) continued to increase asymptotically as K rose from 1 to 10. Therefore, we determined the optimal number of clusters using the metric ΔK (Evanno et al. 2005). Graphs of LnP(D|K) and ΔK are available in the Appendix, and

the posterior value for *r*, the locality parameter, was less than one in every analysis, implying a significant degree of information content in the locality data. The optimal number of clusters for the entire clade A data, which included 360 individuals, was K=2 (Figure 4). This resulted in two clusters with very little admixture that corresponded to 1) all nominal *C. temensis* and 2) *C. pinima* et al. We divided these data and ran the program separately on each set. For *C. temensis* (Adiv1), ΔK indicated that K=2 was optimal, but we saw a gradient in admixture from one cluster to the other, indicating that a single cluster (population), potentially exhibiting isolation by distance from the north to south, was a better explanation of the data. For *C. pinima* et al. (Adiv2), ΔK indicated that K=3 was optimal. Of these three clusters, two (Adiv2-1 and Adiv2-2) included *C. pinima*, *C. jariina*, *C. thyrorus*, and *C. vazzoleri* with a significant degree of admixture between the two clusters, while the third cluster (Adiv2-3) included *C. piquiti*, *C. melaniae*, and *C. mirianae*. Two localities of nominal *C. pinima* exhibited a significant degree of assignment to this third cluster: Tocantins and Paru. The Tocantins *C. pinima* were observed in the mtDNA tree to exhibit *C. piquiti* haplotypes, which suggests this admixture results from introgressive hybridization (*C. piquiti* are found adjacently). For the Paru *C. pinima*, there was no evidence of haplotype sharing in the mtDNA tree. Further, these fishes are not adjacent to a locality where *C. mirianae*, *C. melaniae*, or *C. piquiti* are found, and intervening localities show no evidence of admixture. Looking at the data more closely, it appears that the alleles that are exhibited in common between the Paru *C. pinima* and the Suia Missu *C. mirianae* (the most similar non-*C. pinima* locality) are also found in low frequency throughout the distribution of nomimal *C. pinima*. Thus it appears that these localities have independently evolved higher frequencies of these alleles, creating an artificial pattern of similarity.

To further understand structuring within these 3 clusters of Adiv2, we removed the samples of (hybrid) *C. pinima* from the Tocantins and analyzed separately the individuals from Adiv2-1+Adiv2-2 and Adiv2-3. For this latter division, LnP(D|K) and ΔK both indicated that K=4 was optimal, which corresponded to *C. melaniae, C piquiti,* and separated the localities for *C. mirianae* that lie in separate Amazonas tributaries (Suia Missu and Alta Floresta) (Figure 4). There was also a smaller mode in ΔK at K=2 that corresponded to *C. piquiti* vs. *C. melaniae*+*C. mirianae*. This indicates that while *C.*

piquiti subpopulations are clearly connected by gene flow, there is insufficient data to estimate connectivity between *C. melaniae* and *C. mirianae*. For Adiv2-1+Adiv2-2, ΔK indicated that K=2 was optimal. Again, there was a significant admixture between these clusters at several localities, indicating that either these localities represent a single species with rather strong and complicated population structure or two species which are hybridizing in several localities. In addition, we observed that the placement of localities in clusters did not correspond to geography in a simple way. For example, the larger (magenta) cluster included most central localities, stretching from the Madeira tributaries (Machado) to the mouth of the Amazon (Araguari) and also included nominal *C. vazzoleri* from Oriximiná. The second cluster (pink) included the Tapajós localities (Jacareacanga, Itaituba) and those in or near the mouth of this river, but also, the non-adjacent *C. vazzoleri* from the Jatapu and *C. jariina* in the Jari. Other localities exhibit a split assignment to the two clusters, including the type localitity for *C. pinima*, the Curuá-Una. This division into two clusters is congruent with the major pattern observed for these species in the mtDNA tree.

Analysis of this clade A dataset with STRUCTURAMA, without division, resulted in eight to nine clusters in the mean partition under a mean expectation of five clusters, nine to ten clusters under a mean expectation of ten clusters, and ten to eleven clusters under a mean expectation of fifteen clusters. As nine clusters was the most frequently observed number of clusters (highest cumulative posterior probability), we inferred this to be optimal for this dataset. However, assignment of individuals in those four runs that inferred nine clusters in the mean partition was not entirely consistent between runs. Therefore, the summary of assignment across runs was made with CLUMPP (Figure 4). Most of these clusters were similar to those found in the above divide-and-reanalyze STRUCTURE analyses. In addition, STRUCTURAMA emphasized the distinctiveness of the Machado and Aripuanã *C. pinima*. However, as mentioned, this analysis did not allow admixture between the clusters; split assignment in this CLUMPP summary represents assignment to different clusters between runs. Analysis of admixture in these data with K=9 using STRUCTURE showed a high degree of admixture between these clusters, in particular between Machado, Aripuanã, and other *C. pinima* localities.

As above, the LnP(D|K) in STRUCTURE for the microsatellite data of clade B species, *C. ocellaris*+*C. orinocensis* et al., continued to increase asymptotically with K, so we used the rate of change between K (ΔK) to estimate optimal clustering. Also as above, the value for the locality parameter, *r*, was always estimated to be less than one. For the full dataset of 666 clade B individuals, the optimal K was K=2. This corresponded to *C. intermedia*+*C. orinocensis* (Bdiv1) and *C. ocellaris*+*C. monoculus*+*C. pleiozona*+*C. kelberi*+*C. nigromaculata* (Bdiv2) (Figure 4). There was some overlap between these two clusters at several localities of the *C. ocellaris* et al., but most of these localities were not sympatric or contiguous with *C. orinocensis* or *C. intermedia*, implying it is probably a result of allele size homoplasy. After dividing the dataset for reanalysis, we found that the optimal clustering for Bdiv1 was K=2, which corresponded to *C. orinocensis* and *C. intermedia*. Upon analyzing *C. orinocensis* separately (Bdiv1-1), K=2 was determined to be the optimal clustering, but we observed a gradient in admixture moving from one end of this species' distribution to the other. As with *C. temensis*, we interpret this to imply that a single cluster is truly optimal, possibly with isolation-by-distance between localities. For *C. intermedia* (Bdiv1-2), we again found that K=2 was optimal for these data, but in contrast to *C. orinocensis*, this clustering distinguished one subpopulation of *C. intermedia* from the rest (Caura). As the mtDNA tree implied all of these individuals were very closely related, we did not further subdivide *C. intermedia* for analysis. For *C. ocellaris* et al. (Bdiv2), we observed that the optimal number of clusters was K=2. This emphasized the distinctness of several localities in the Negro and Orinoco Rivers (all nominal *C. nigromaculata* and several *C. monoculus* localities) relative to the remainder. However, a number of other localities also showed a significant degree of admixture between these clusters. We, therefore, did not divide and reanalyze these data.

Analysis of the clade B data, *C. ocellaris*+*C. orinocensis* et al., with STRUCTURAMA resulted in nine clusters with the highest cumulative posterior probability. Analysis with a mean expectation of five clusters resulted in seven to nine clusters with a posterior probability greater than zero, an expectation of ten clusters resulted in eight to twelve clusters, and an expectation of fifteen clusters resulted in nine to thirteen clusters. However, unlike with *C. temensis* et al., analysis of the *C.*

ocellaris+*C. orinocensis* et al. dataset did not support a single number of clusters in each run; rather, each run exhibited a divided posterior probability for two to three numbers of clusters. For example, with an expectation of five, one run supported K=7: 0.07, K=8: 0.37, K=9: 0.56, with nine clusters in the mean partition. Summing the posterior probability across all twelve runs, nine clusters received the highest cumulative posterior probability, followed by twelve. For those five runs that exhibited nine clusters in the mean partition, we summarized individual assignment to cluster across runs with CLUMPP (Figure 4). As above, because this analysis did not allow admixture, split assignment in this CLUMPP summary represents assignment to different clusters between runs. In all STRUCTURAMA runs, *C. orinocensis* and *C. intermedia* were separated from *C. ocellaris* et al. In addition, this analysis emphasized the distinctiveness of several sets of localities in the *C. ocellaris* et al. group (Bdiv2) in addition to those in the Negro and Orinoco: both localities of nominal *C. pleiozona* (Guajará-Mirim and Abunã) plus two localities of *C. monoculus* in the middle Madeira (Humaita and Cunia); both localities of nominal *C. kelberi* (São Felix and Tocantins); *C. ocellaris* from the Maroni; *C. ocellaris* from the Cuyuni; and *C. monoculus* from the Tapajos (Itaituba and Jacareacanga). However, several of these were assigned to different clusters between runs. Further, when these data were analyzed with Structure with K=9, some of the same localities were emphasized (e.g. Guajará-Mirim et al., São Felix et al.), while others were lumped together (e.g. Itaituba+Jacareacanga), and still others were emphasized as distinct. Moreover, there was a significant degree of admixture between these clusters. Taken together, these STRUCTURE and STRUCTURAMA results portray localities of *C. ocellaris* et al. (Bdiv2) along the main Amazonas as more homogenous (nominal *C. monoculus*), while localities farther away from center, in the tributaries and satellite drainages, are more distinct (nominal *C. pleiozona*, *C. nigromaculata*, *C. ocellaris*, *C. kelberi*, and some *C. monoculus*). However, there was inconsistent clustering of localities into separate meta-populations (species), implying that one cluster is the best explanation for these five putative species.

Microsatellite analysis of hybridization

Based on the mismatch between morphology (species ID) and mtDNA or nuclear gene lineages, we identified at least seven instances of introgressive hybridization between species. Analysis of five of these putative hybrid subpopulations using STRUCTURE, along with putatively non-hybrid individuals, showed a range of admixture (Figure 5). One-sample *t*-tests showed that some putative hybrids exhibited no significant admixture (cluster posterior > 0.95; e.g. *C. nigromaculata* at Mavaca), while others exhibited introgression of nearly half their alleles (e.g. *C. temensis* in Guri Reservoir). In the case of the nominal *C. pinima* from the Tocantins, hybrid fishes were more like *C. piquiti* than *C. pinima*! Some failed tests may have been affected by sample size (e.g. *C. intermedia*, Parguaza). Others were significant only until additional non-hybrid localities were added (e.g. *C. orinocensis*, Parguaza), implying that population structure could affect these tests.

Discussion

Species delimitation using molecular data

There is a growing consensus among evolutionary biologists and systematists that species should be treated as hypotheses that are subject to revision in light of data from natural populations (Sites and Crandall 1997; Templeton 2001). Molecular data represent a useful resource in this context because they provide an assessment of effective genetic exchange between groups of individuals that are hypothesized to constitute an evolving biological entity (Templeton 2001). Any set of data used to infer species boundaries, however, will suffer from the well-known systematic adage that distinctiveness of two species will be the inverse of the number of specimens examined. In effect, this implies that to adequately test species hypotheses, it is necessary to sample densely-enough, and adaptively, in a manner designed to test observed discontinuities between putative species (Morando et al. 2003). In addition, the use of independent data sources (morphology, mtDNA, microsatellites, etc.) provides a more robust test of species hypotheses since any one data source may provide a misleading estimate of cohesiveness or disjunction. Nevertheless, species, particularly widespread species, are often contentious to delimit, stemming from the ambiguous correspondence between a species taxon and the

biological entity to which it is meant to refer (Hey et al. 2003). Our estimate of species, like any other, is inherently subjective and a product of our own evolutionary paradigms and inclinations. In interpreting our data regarding species taxa, we emphasized groups that would likely be subject to the same ecological and demographic constraints, and whose zones of intergradation (hybridization) appeared to restrict the overall exchange of these forms (Templeton 1989; Mallet 2007b).

We interpret our present results to support the discrimination of eight rather than fifteen species in the genus *Cichla*: *C. orinocensis*, *C. intermedia*, *C. ocellaris*, *C. temensis*, *C. melaniae*, *C. mirianae*, *C. piquiti*, and *C. pinima*. The remaining species appear to form species complexes within these taxa rather than independent biological entities. Specifically, we suggest that the nominal species *C. monoculus*, *C. pleiozona*, *C. nigromaculata*, and *C. kelberi* be considered subspecies or evolutionarily significant units (ESUs) of *C. ocellaris sensu lato*. As *C. ocellaris* Schneider, 1801 was the first described species of *Cichla*, this name should apply based on the rules of precedence. Similarly, we believe that *C. jariina*, *C. vazzoleri*, and *C. thyrorus* are better considered sub-specific designations of *C. pinima sensu lato*. However, in this case these taxa were all described in a single review, so the nomenclatural rules are ambiguous. However, as the nominal species *C. pinima* appears to show less incongruence given current results, we suggest that this name be used to refer to this species group.

Each of these delimited species is easily distinguishable based on morphology from the other delimited species (Figure 1b; see also Kullander and Ferreira 2006). Further, these species showed exclusive lineages of mtDNA and separate clusters in the microsatellite analysis, implying that they experience, and have experienced in the past, more exclusive gene flow than with heterospecifics. For example, while *C. temensis* and *C. orinocensis* exhibited an optimal number of clusters of K=2, many individuals were admixed between these clusters. Further, both of these species showed mtDNA lineages that were exclusive to these species (notwithstanding hybridization) and distributed heterogeneously throughout their range. While *C. orinocensis* had two clades, these were found together in one geographically intermediate locality, and there was no congruence between the transition between mtDNA clades and microsatellite clusters geographically. Moreover, for both *C. temensis* and *C. orinocensis*, within each of these mtDNA lineages,

haplotypes were distinguished by few mutations compared to hetero-specific haplotypes, suggesting a much more recent coalescence. Similarly, *C. intermedia* exhibited two optimal clusters for its microsatellite data, but unlike in *C. orinocensis* or *C. temensis*, one of these clusters corresponded to a single locality: the Caura. Upon examining the data, this locality exhibited a reduced diversity of alleles in comparison to remaining localities but few unique alleles. Further, the recent coalescence of mtDNA between these subpopulations implies a recent separation, perhaps followed by a bottleneck event.

Cichla piquiti and *C. melaniae* were each identified as single cluster in the microsatellite analyses, while *C. mirianae* was suggested to have two clusters corresponding to its subpopulations in the Xingu and Tapajós drainges. While *C. piquiti* has an mtDNA lineage that is well-differentiated from other species, *C. melaniae* and *C. mirianae* had mtDNA haplotypes that were exclusive but more similar than amongst other delimited species. There was also some ambiguity in the microsatellite results as to whether these latter two species corresponded to one cluster rather than three (see graphs of LnP(D|K) and ΔK in the Appendix). Given our current sampling design, we could not reject the hypothesis that there were two biological entities, but we suggest that a denser sampling in the middle Xingu and upper Tapajós be done to further test this hypothesis.

The mtDNA of *C. pinima* was divided into two well-differentiated clades that, while largely allopatric, showed a complex geographical pattern and which were found together in two localities. The mtDNA of *C. vazzoleri*, *C. thyrorus*, and *C. jariina* was subsumed within these two clades. Similarly, the microsatellite data for these nominal species was best divided into two overlapping clusters. While there was large congruence between these two datasets, it was not strict. For instance, while the nominal *C. vazzoleri* from the Jatapu were clustered with the Tapajós *C. pinima* and *C. jariina* based on microsatellies, their mtDNA resided in the clade containing the central (mainstem) *C. pinima*, Oriximiná *C. vazzoleri*, and *C. thyrorus*. The inference of two rather than four clusters and the complex geographical structure among them suggests that the described species have shared gene flow too recently to be evolving separately. Whether these two clusters represent a single species with a very complex population structure or two species with one or more zones of hybridization is unclear from the current data, but may be addressed through the use of coalescent-based models (Knowles and Carstens 2007;

Carstens and Dewey 2010). In any event, these species were originally distinguished on the basis of very subtle difference in color pattern and overlapping meristic and morphometric data, and considering the present data, it seems more consistent to consider them subspecies or ESUs of a more inclusive species (*C. pinima sensu lato*).

Similarly, for the species in clade B1, discrimination was based on very subtle difference in color pattern and overlapping meristic and morphometric data. However, several of these putative species exhibited unique mtDNA lineages. Nevertheless, the microsatellite data did not distinguish these groups of sub-populations as being more dissimilar from each other (i.e. having a more exclusive history of gene flow) than some sets of subpopulations within nominal *C. monoculus*. Moreover, there were several sets of subpopulations that showed incongruence between microsatellite affinity and mtDNA clade. For example, while localities in the upper Purus River and middle and lower Madeira River exhibited mtDNA more similar to the *C. pleiozona* clade, they always grouped with *C. monoculus* based on microsatellites. Similarly, while middle Tapajós (Jacareacanga) fishes had mtDNA more closely related to *C. pleiozona*, these fishes appeared more similar to *C. kelberi* with microsatellites. This suggests that despite having mtDNA lineages that coalesce rather deeply, there is little evidence for reproductive isolation between even the most divergent lineages. Further, based on the mtDNA genealogy, it appears that the populations bearing the more mtDNA divergent lineages, *C. pleiozona*, have been exchanging genes at a low rate with the central populations (*C. monoculus*) at least as long as *C. kelberi* mtDNA has been coalescing independently. While it is evident that some of these subpopulation groups show unique characteristics that imply a reduced rate of gene flow, some of the geographically-restricted nominal species (e.g. *C. ocellaris* from the Essequibo and Pirara, *C. kelberi* from the Tocantins and São Felix do Araguaia) appear no more differentiated than subpopulations within more widespread nominal species (e.g. *C. monoculus* from Itaituba). Rather, we interpret the data to imply that meta-populations in this clade form a widespread genetic mosaic. Under this interpretation, small groups of subpopulations experience a gradient of gene flow with other such groups. At the high end, in this case closer to the main Amazonas channel, nuclear homogenization is common and mtDNA, while often developing unique clades, is exchanged occasionally among groups. At the low end, farther away from the

Amazonas, gene flow is low enough (or isolation long enough) for divergence in mtDNA and dissimilarity in microsatellite patterns. However, with the changes in river drainage through time (i.e. geodispersal), homogenization occurs between even more disparate subpopulation groups. Why these subpopulations do not develop reproductive isolation is a question that should be addressed with directed study. Perhaps the slow rate of molecular evolution in *Cichla* and the constraints of piscivorous foraging strategy limit opportunities for divergence (Collar et al. 2009). In any event, the observation that these populations appear demographically interchangeable and freely interfertile suggests that they should be considered subspecies or ESUs rather than separate species.

Our results differed from the morphological review of Kullander and Ferreira (2006), but we do not think the incongruence stems from different data sources *per se*. As we alluded to above, the morphological characters upon which the species in these complexes were discriminated seem of questionable utility to us. Further, we have examined an overall greater number of specimens, with the intention of testing apparent disjunctions with targeted sampling. It is important to point out, of course, that in the course of this study we had the advantage of reviewing the results of Kullander and Ferreira (2006) and collecting data with an explicit intent to test their hypotheses. On the other hand, there are a number of reasons why our dataset could be incorrect regarding species boundaries. We would be remiss not to point out that our sample sizes for some species were rather low, particularly for *C. thyrorus* (N=2) and *C. jariina* (N=9), each from a single locality. This may explain why these taxa were not found to be more distinct. On the other hand, *C. ocellaris* from the Maroni (N=2) and Cuyuni (N=1) were emphasized as distinct in the STRUCTURAMA analysis despite low sample size. Further, while the STRUCTURE/STRUCTURAMA analyses might be misled by this effect, low sample size would not explain the sharing of mtDNA haplotypes between nominal taxa. Another reason for incongruence could be the mutational constraint or bias of unknown strength that microsatellite loci are suspected to exhibit (Ellegren et al. 1995; Rubinsztein et al. 1995). This may, for instance, explain why some groups of clade B1 species did not appear more dissimilar despite having distinct mtDNA lineages. Alternatively, this could have caused the Negro localities of *C. nigromaculata* and *C. monoculus* to appear overly dissimilar, leading us to end our division and reanalysis with STRUCTURE prematurely.

However, while there may be unknown biases in some populations, our observation of a significant degree of size variation and number of alleles in each species group implies that there should be no lack of power with these loci. Moreover, while any one locus can provide a misleading estimate of population structure, our use of multiple loci and different analytical methods allowed us to estimate species limits while taking into account the idiosyncracies of each data set. For example, bias in the microsatellites would not explain why so many populations of nominal *C. monoculus* shared mtDNA lineages with *C. pleiozona* in adjacent localities.

We chose to refer some nominal species to a sub-specific category such as subspecies or ESU, while some would argue that any evidence available should be used to elevate populations to the species rank. One common reason is in order to make evolutionarily unique population more likely to be the focus of conservation efforts. While a laudable goal, we feel this is misguided. We agree with Hey et al. (2003) that this priority will tend to shift the focus of taxonomic decisions away from biological phenomena and towards political criteria (see also Karl and Bowen 1999). We also expect that if systematists choose to enter the political arena in this manner, eventually policy makers and resource managers will come to suspect their objectivity. Rather, it seems likely that the current legislative focus on species (e.g. 'endangered species') rather than less inclusive but evolutionarily unique forms should be brought more in line with current biological paradigms (Moritz 1994; Mallet 2007b; but see recent conservation strategies for Pacific salmon, Hey et al. 2003). We find that our nomination of these taxa for subspecific categories is more consistent with their evolutionary history and evidence of reproductive isolation.

We analyzed the microsatellite dataset by iteratively dividing the data and reanalyzing with STRUCTURE, and we found these results to be more directly informative than those from STRUCTURAMA. In contrast, a recent review of multilocus analyses of species boundaries using clustering approaches found that STRUCTURAMA outperformed STRUCTURE when the optimal number of clusters were chosen in STRUCTURE using LnP(D|K) or ΔK (Hausdorf and Hennig 2010). It was unclear from their paper, however, how they constructed their analyses for each program, and how they interpreted their STRUCTURE/STRUCTURAMA results with respect to species

hypotheses. We interpreted our divide and reanalyze protocol to be more consistent with the findings of Evanno et al. (2005), where ΔK was observed to identify the highest hierarchical level of structure in the data. Were one to interpret the first optimal clustering identified by ΔK as the best estimate of species, it is easy to see why this analysis could fail to identify species-level meta-populations. On the other hand, while the ability to estimate the number of clusters stochastically with STRUCTURAMA is appealing, we found that analytical constructions without admixture may overestimate the number of clusters or provide ambiguous results. In the present case, with a dataset as large and complex as ours, we found it computationally intractable to implement the admixture model in our analyses with STRUCTURAMA. However, when we examined the data in STRUCTURE using the number of clusters estimated using STRUCTURAMA, we found that there was a high degree of overlap between many clusters, supporting the notion that using the admixture function may be beneficial for estimating the number of clusters in complex datasets.

An alternative method of analysis for our molecular data would be fully parameterized models such as those implemented in IM (Nielsen 1998; Hey and Nielsen 2007) or LAMARC (Kuhner 2006). These methods use coalescent simulations to estimate population sizes and migration rates between localities. These models have the advantage of integrating multiple data types (e.g. microsatellites and mtDNA), implementing more realistic models for the mutation of microsatellites (e.g. Brownian motion), and integrating across uncertainty in parameter estimates. Two limitations prevented us from utilizing these models here. First, it would be difficult to ensure convergence and accurately estimate parameters ranges for all of the morphotypes by localities by loci included in our dataset with available computational and time resources. Secondly, these models make stricter assumptions about stable mutation rates of the loci included since the simulated coalescent trees are inherently ultrametric. Thus, while these models could be useful for analyzing subsets of our dataset, we did not find them practical with the overall dataset.

Introgressive hybridization in Cichla

Another benefit of collecting large amounts of molecular data, as we have done, is the opportunity to investigate the frequency and genomic extent of introgressive hybridization between our delimited taxa. Introgression is a special class of hybridization in which multiple generations of backcrossing between hybrid and non-hybrid individuals allows the movement or fixation of heterospecific alleles in a population. Based largely on mtDNA, and to a smaller extent on morphology (species ID) and nuclear gene genealogies, we identified seven putative instances of recent introgression between species. Overall, these constituted a very small proportion of the overall number of individuals in our dataset (<5%). On the other hand, this hybridization involved 6 of our 8 delimited species of *Cichla*, or 9 of the 15 described species. This included hybridization between sister species (*C. orinocensis* x *C. intermedia*) and more distantly related species (*C. temensis* x *C. orinocensis* and *C. ocellaris s.l.*). The observation of apparent viability of hybrids between even more divergent species suggests that reproductive isolating mechanisms, where they exist, are likely to be pre-zygotic. Considering the divergence in color pattern between our delimited species, this, perhaps coupled with some behavioral cues, would seem like an effective method of mate choice for these visual predators.

These observations and conjectures should stimulate investigation as to what conditions lead to the break down of reproductive isolation among *Cichla* species. During our (unstandardized) collections, we observed that, in those localities unaltered by humans where we later inferred introgression, individuals of the recipient species, if not both recipient and donor, were relatively rare, perhaps implying that hybridization was facilitated by mate scarcity (Arnold 1997; Dowling and Secor 1997; Randler 2001; Grant et al. 2005). For instance, *C. nigromaculata* is relatively rare in the upper Orinoco and Casiquiare region, while *C. temensis* is more common. This may have facilitated the founding of a subpopulation that was fixed for *C. temensis* mtDNA, but was morphologically *C. nigromaculata* (*C. ocellaris s.l.*). Several of our other inferred instances of hybridization were from habitats altered by human influence. For instance, the *C. temensis* x *C. orinocensis* from Guri Reservoir experience genuine lacustrine conditions, a rare challenge for Neotropical fishes. Similarly, *C. pinima* x *C. piquiti* from

the Tocantins were collected downstream of the Tucuruí Reservoir, a region which has experienced a radically different flow regime since the erection of the dam (based on conversations with local residents). Hybridization was also reported for fishes in the Balbina Reservoir on the Uatumã River (Brinn et al. 2004). The Cuyuni River, on the other hand, where we inferred hybridization between *C. ocellaris* and *C. temensis*, has shown a greatly increased sedimentation over several decades due to dredging for gold (Willis, pers. obs.). Interestingly, for the Guri and Cuyuni fishes, our observation of morphology indicated that something was peculiar about these fishes, while all other putative hybrids conformed to the morphology of non-hybrid parental individuals.

As most hybrids were identified here using mtDNA, we were concerned that mtDNA would provide a biased estimate of the extent of introgression in *Cichla*. It has been suggested in recent years that mtDNA may introgress more readily than nuclear genes (Chan and Levin 2005), perhaps inflating the apparent impacts of hybridization, although there are clear instances of the opposite phenomenon (Di Candia and Routman 2007). Where sample size permitted, the focused microsatellite analyses using STRUCTURE showed that introgression in the nuclear genome ranged from extensive to negligible (Figure 5). This suggests that the forces governing introgression may be different in each case. However, even if the only lasting indicator of introgression is mtDNA (e.g. the two mtDNA clades of *C. orinocensis*), this nevertheless shows that 1) early hybrids were viable and fertile and 2) subsequent backcrossing occurred. These points imply that opportunities for "adaptive introgression", the transfer of adaptive mutations and an increase in genetic diversity not constrained by *in situ* mutation (Anderson and Stebbins 1954; Harrison and Rand 1989; Buerkle et al. 2000), are more common than is traditionally assumed. Particularly in species that are constrained by a slow rate of mutation, such as appears to be true of *Cichla* (Farias et al. 1999; López-Fernández et al. 2005), introgression may increase the genetic diversity and adaptive potential of a species and even stimulate lineage diversification (Anderson and Stebbins 1954; Lewontin and Birch 1966; Mallet 2005, 2007a; Stelkens and Seehausen 2009).

Conclusions

We applied a qualitative method of observing the congruence between markers for genetic exchange and disjunction between putative species groups in order to delimit species in this widespread and morphologically conservative genus. Instead of following a strict criterion to identify species boundaries, we examined the full range of genetic overlap among subpopulations to distinguish gene flow from introgressive hybridization based on observed discontinuities. No doubt some readers will interpret our method as being overly subjective, but we counter that the inherently ambiguous nature of species groups makes any objective delimitation of species limited in utility either taxonomically or as a framework for further evolutionary study. Rather than stress the species groups we have delimited *per se*, we emphasize the genetic overlap between our delimited species and among their contained populations, and call attention to the evolutionary processes that overlap implies.

Based on extensive sampling and multilocus analyses, we concluded that *Cichla* contains fewer species that correspond clearly to biological entities than are currently recognized. While at least two of these species contain evolutionarily significant units that are in need of conservation, these populations did not appear to be distinct in their mtDNA, microsatellites, or both. While some Neotropical fish species groups exhibit a smaller or more fragmented geographic range than *Cichla*, and could thus be expected to show a higher degree of microendemism, many are widespread and may experience long-distance genetic exchange over evolutionary time. We suggest that systematists focusing on widespread Neotropical fishes work to test apparent morphological or molecular disjunctions before erecting specific categories. Intriguingly, while we estimated fewer cohesive biological entities than expected, we inferred introgressive hybridization at a higher than expected rate. Although the extent of introgression varied across cases, its frequency suggests that hybridization may play a hitherto unanticipated but important role in the adaptation and/or diversification of Neotropical freshwater fishes and other tropical lineages.

Table 1. Sampling localities and sample sizes for mtDNA / nuclear sequences / microsatellites.

	Locality	*temensis (t)*	*pinima (p)*	*vazzoleri (v)*	*thyrorus (y)*	*jariina (j)*	*piquiti (q)*	*melaniae (a)*	*mirianae ®*	*intermedia (i)*	*orinocensis (o)*	*ocellaris (c)*	*monoculus (m)*	*nigromaculata (n)*	*kelberi (k)*	*pleiozona (z)*
TI	Tigre										8/3/10					
GU	Guanipa										4/2/4					
BJ	Buja										2/-/2					
GR	Guri Reservoir (Caroni)	11/1/8														
SI	Sipao	10/-/10									10/5/10					
CA	Caura									15/1/12						
CV	Cunavichito	1/-/1									3/-/3					
CP	Capanaparo	10/-/10									10/2/10					
CI	Cinaruco	12/1/26								10/1/10	11/3/11					
PZ	Parguaza	2/-/2								2/2/2	12/-/12					
AT	Atabapo	10/1/10								2/-/2	10/1/10					
VE	Ventuari	9/-/9								12/1/12	12/2/25					
OR	Orinoco									4/-/4						
IG	Iguapo	1/1/-								10/-/9						
OC	Ocamo									10/1/9						
MV	Mavaca													10/2/10		
CR	Curamoni	3/-/3									7/-/7					
PA	Perro de Agua										16/-/16					
CQ	Casiquiare	1/-/1								15/1/15	2/-/2					
PS	Pasiba	10/1/10									17/4/17			10/1/9		
UA	Uaupes	11/-/10									20/-/20					
IM	Ia-mirim	1/-/-									5/-/5					
TE	Teá	1/-/1									10/-/10			1/-/-		
MR	Marauiá	9/-/7									10/-/10			11/1/10		
UE	Uneiuxi	10/-/5									23/4/10					
DA	Daraá										2/-/2					
PT	Preto	1/-/1														
BC	Barcelos	7/-/7									1/-/1		10/4/10			
PI	Pirara (Takutu)	5/1/5										11/1/11				
ES	Essequibo (Rupununi)											13/3/1(				
CY	Cuyuni											1/1/1				
MA	Maroni											2/2/2				
XE	Xeruini	9/2/9									10/6/10					
TA	Tapera	12/2/10									10/5/10					
UN	Unini	16/3/10									12/4/12		12/3/10			
NA	Novo Airão	11/-/9									10/-/10		11/-/10			
PE	Preta da Eva	4/-/4									6/-/6		2/-/2			
UR	Urubu	3/-/3														
PL	Pucallpa												2/-/-			
IQ	Iquitos												8/2/4			
TB	Tabatinga												22/-/9			
JA	Juruá (Carauari)												8/-/-			
EI	Eirunepé												3/-/3			
CS	Cruzeiro do Sul												10/-/10			
AM	Lago Amaná												10/-/9			
TF	Tefé												8/-/7			
CO	Coari												6/-/6			
PP	Piagaçu-Purus												10/-/10			
TP	Tapauá												11/-/10			
LB	Labrea												2/-/2			
BA	Boca do Acre												20/-/10			
MC	Manacapuru												10/-/10			
IA	Igapo-Açu	10/2/10											10/1/10			
BO	Borba												10/-/10			
CM	Canumã		10/1/10										13/-/13			
AP	Aripuanã		13/2/10										3/-/3			
HU	Humaita												9/-/9			
MD	Machado		2/1/2													
CN	Cunia												20/-/10			
CC	Canaçari												10/-/8			
MS	Maués		9/1/8										10/-/7			
JT	Jatapu (Uatumã)			5/5/5												

Table 1. *continued*

	Locality	*temensis (t)*	*pinima (p)*	*vazzoleri (v)*	*thyrorus (y)*	*jarina (j)*	*piquiti (q)*	*melaniae (a)*	*mirianae ®*	*intermedia (i)*	*orinocensis (o)*	*ocellaris (c)*	*monoculus (m)*	*nigromaculata (n)*	*kelberi (k)*	*pleiozona (z)*
NH	Nhamunda		10/-/10										10/-/9			
TS	Terra Santa												10/-/10			
TR	Trombetas (abv. rapids)				2/2/2											
OX	Oriximiná			15/2/14												
LG	Lago Grande		9/-/9													
TL	Tapajós mouth		10/2/10													
IT	Itaituba		10/3/9										14/-/9			
JC	Jacareacanga		8/-/8										5/-/5			
CU	Curuá-Una		5/1/5													
PU	Paru		6/1/6										13/-/10			
GA	Guajara		10/-/9													
VX	Vittoria do Xingu		4/1/4										10/-/10			
JR	Jari (lower)												10/1/10			
JU	Jari (above waterfalls)					9/5/9										
AR	Araguari		6/1/6										2/1/2			
AF	Alta Floresta								5/5/5							
SM	Suia Missu								10/2/10							
XA	Xingu (Altamira)							2/2/2								
IR	Iriri							19/5/13								
TO	Tocantins (Baião)		4/-/4												10/-/10	
AG	Araguatins						10/-/7									
SF	São Felix do Araguaia						10/5/10								10/5/10	
AB	Abunã															7/-/7
GM	Guajará-Mirim															10/4/9
MP	Manuripi															12/-/-
YT	Yata															9/-/-
SC	Secure															2/-/-
SN	San Martin															6/-/-
IC	Ichilo															4/-/-
PG	Paragua															8/-/-
Totals:		190/15/181	116/14/110	20/7/19	2/2/2	9/5/9	20/5/17	21/7/15	15/7/15	80/7/75	243/41/245	27/7/24	324/12/257	32/4/29	20/5/20	58/4/16

Table 2. Allele diversity and size range for the microsatellite loci.

	Clade A		*Clade B1*		*Clade B2*	
locus	num. alleles	size range	num. alleles	size range	num. alleles	size range
Cint22	27	127-185	33	121-195	36	129-203
CoriA6	22	255-309	18	257-289	23	257-333
CoriB3	20	201-241	20	189-231	10	191-221
CoriB6.2	29	268-338	44	253-334	6	266-274
CoriD12	12	150-174	26	148-198	10	152-170
CoriF12	36	254-328	35	228-358	29	228-290
CoriG4	19	286-326	7	286-306	10	276-322
CpinC1	19	221-259	7	219-241	5	219-227
CpinC11	18	219-257	20	205-247	23	207-261
CpinD2	36	267-325	35	267-323	32	273-351
CpinE3	30	260-324	34	270-338	36	274-354
CSM2	33	230-278	26	221-258	23	219-261

Figure Legends

Figure 1. a) Map of sampling localities, with species collected in those localities. Codes for localities and species follow Table 1 and Figure1B. b) Representative color patterns for the 15 described species of *Cichla*.

Figure 2. Mitochondrial genealogy (ML phylogram) inferred from TREEFINDER, with haplotypes as terminals. a) Full phylogram, with major clades identified. In B, C, and D, localities follow Table 1, branch values are bootstrap percentages, and terminals with asterisks (*) included both mtCR and mtATP (see text). b) Clade B2. c) Clade A. d) Clade B1.

Figure 3. ML genealogies for the nuclear genes Mitf (a) and Xsrc (b), with haplotypes as terminals, inferred using TREEFINDER. For each haplotype, the number of alleles observed in each described species is listed.

Figure 4. Results of the microsatellite analyses, using the STRUCTURE divide-and-reanalyze approach, STRUCTURE using K (no. clusters) chosen in STRUCTURAMA, and results of the STRUCTURAMA runs with highest posterior K. In each case, figures represent CLUMPP summaries of multiple runs (see text). Species and locality codes follow Table 1. Bold lines indicate where data was divided for separate analysis. a) Clade A. b) Clade B.

Figure 5. Analysis of hybrid localities using STRUCTURE, where allele frequencies were estimated only from non-hybrid individuals. H above a localities denotes putative hybrids based on morphology, mtDNA, and/or nuclear sequences. * denotes mean assignment to parental species significantly less than 0.95; ** denotes significantly less than 0.9. Species and locality codes follow Table 1.

Figure 1

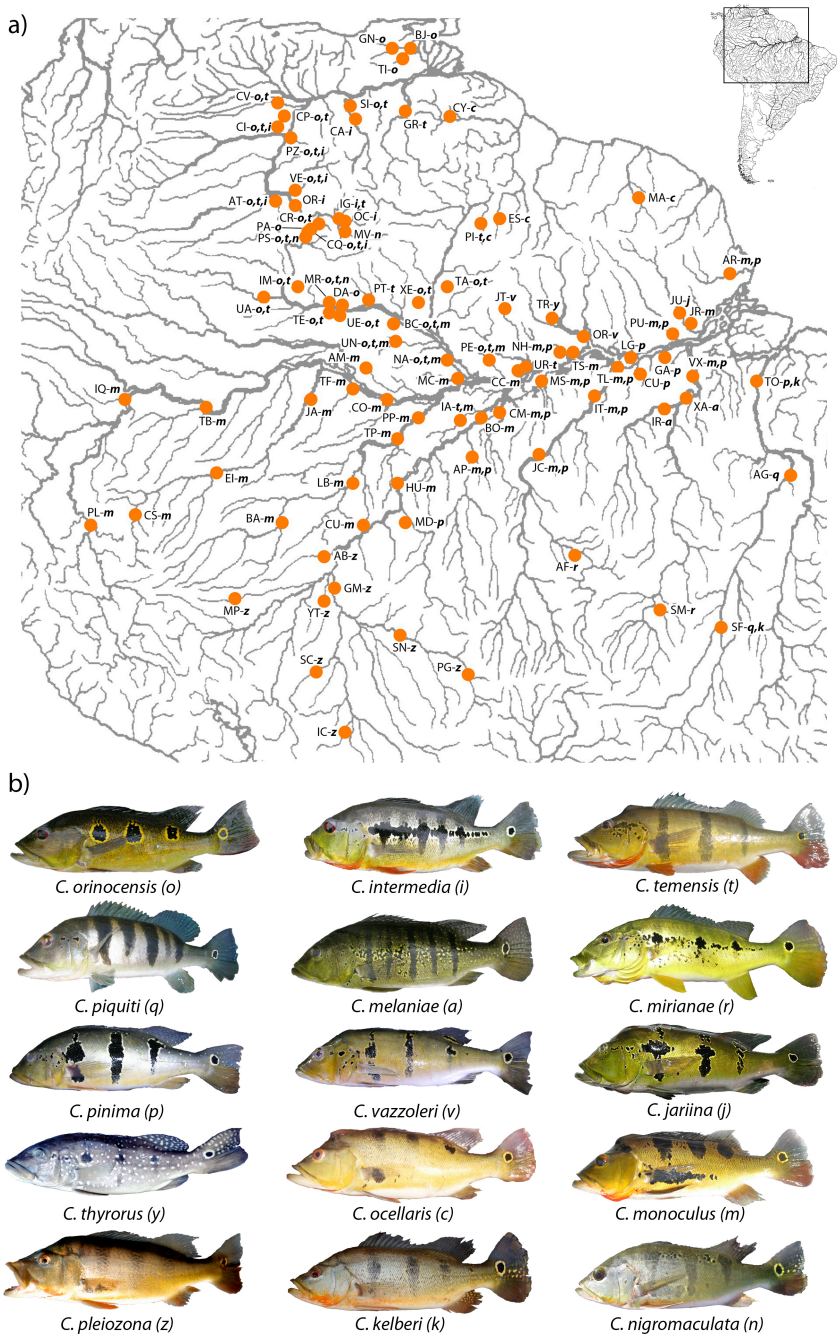

a)
GN-o
BJ-o
TI-o
CV-o,t
SI-o,t
CP-o,t
CY-c
CI-o,t,i
CA-i
GR-t
PZ-o,t,i
VE-o,t,i
AT-o,t,i
OR-i
IG-i,t
MA-c
CR-o,t
OC-i
ES-c
PA-o
MV-n
PI-t,c
PS-o,t,n
CQ-o,t,i
AR-m,p
IM-o,t
MR-o,t,n
DA-o
PT-t
XE-o,t
TA-o,t
JT-v
JU-j
UA-o,t
TE-o,t
UE-o,t
BC-o,t,m
TR-y
JR-m
PU-m,p
UN-o,t,m
OR-v
PE-o,t,m
NH-m,p
LG-p
AM-m
NA-o,t,m
UR-t
TS-m
VX-m,p
TF-m
MC-m
CC-m
MS-m,p
TL-m,p
GA-p
CU-p
TO-p,k
IQ-m
JA-m
CO-m
IA-t,m
IT-m,p
XA-a
TB-m
PP-m
CM-m,p
IR-a
BO-m
TP-m
JC-m,p
EI-m
AP-m,p
AG-q
LB-m
HU-m
PL-m
CS-m
BA-m
CU-m
MD-p
AB-z
AF-r
GM-z
MP-z
YT-z
SM-r
SF-q,k
SN-z
SC-z
PG-z
IC-z
b)
C. orinocensis (o)
C. intermedia (i)
C. temensis (t)
C. piquiti (q)
C. melaniae (a)
C. mirianae (r)
C. pinima (p)
C. vazzoleri (v)
C. jariina (j)
C. thyrorus (y)
C. ocellaris (c)
C. monoculus (m)
C. pleiozona (z)
C. kelberi (k)
C. nigromaculata (n)

Figure 2

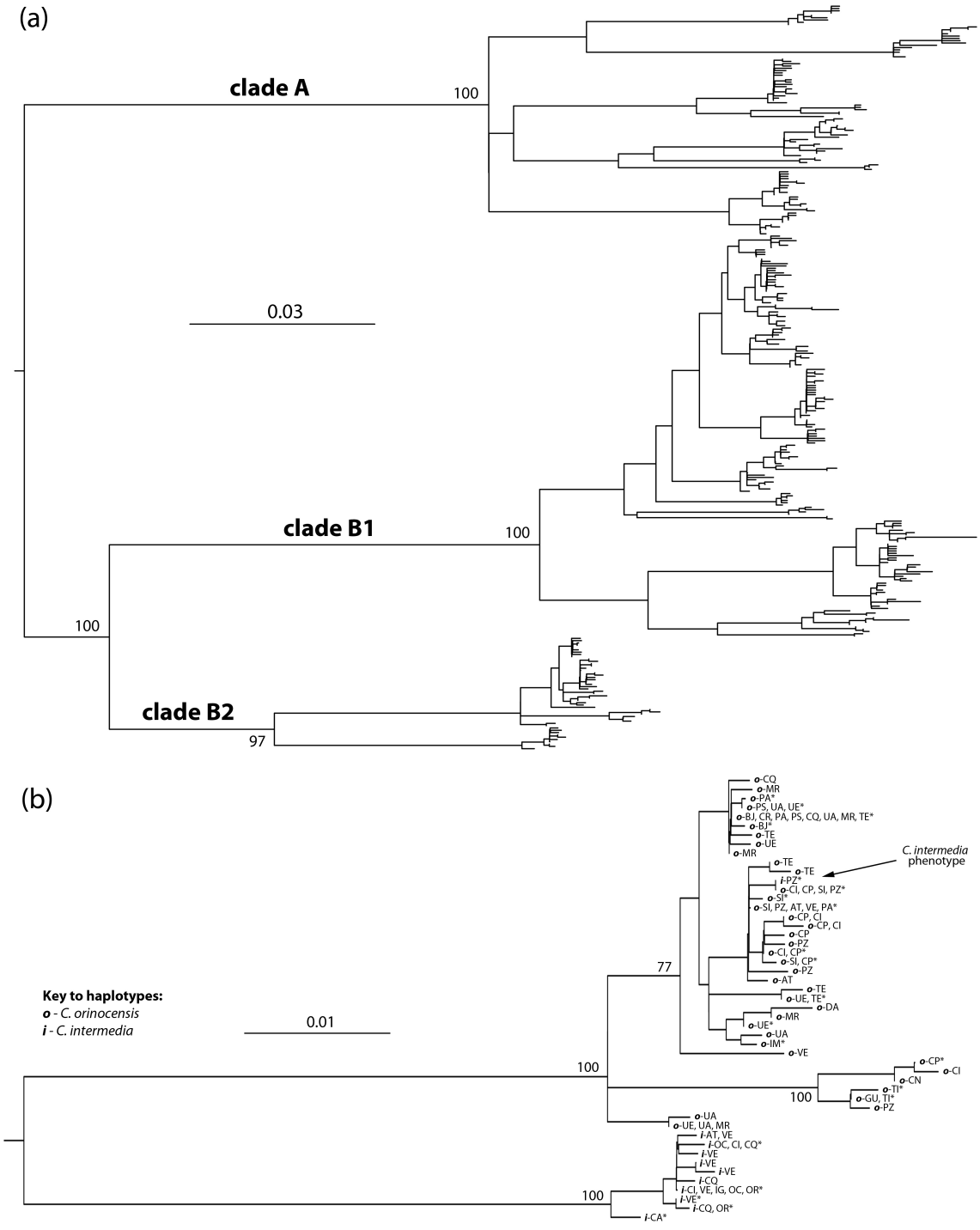

(a)
clade A
100
0.03
clade B1
100
100
clade B2
97
(b)
o-CQ
o-MR
o-PA*
o-PS, UA, UE*
o-BJ, CR, PA, PS, CQ, UA, MR, TE*
o-BJ*
o-TE
o-UE
o-MR
o-TE
o-TE
C. intermedia phenotype
i-PZ*
o-CI, CP, SI, PZ*
o-SI*
o-SI, PZ, AT, VE, PA*
o-CP, CI
o-CP, CI
o-CP
o-PZ
o-CI, CP*
o-SI, CP*
o-PZ
o-AT
77
o-TE
o-UE, TE*
o-DA
o-MR
o-UE*
o-UA
o-IM*
o-VE
Key to haplotypes:
o - C. orinocensis
i - C. intermedia
0.01
100
o-CP*
o-CI
o-CN
100
o-TI*
o-GU, TI*
o-PZ
o-UA
o-UE, UA, MR
i-AT, VE
i-OC, CI, CQ*
i-VE
i-VE
i-VE
i-CQ
100
i-CI, VE, IG, OC, OR*
i-VE*
i-CQ, OR*
i-CA*

Figure 2 *continued*

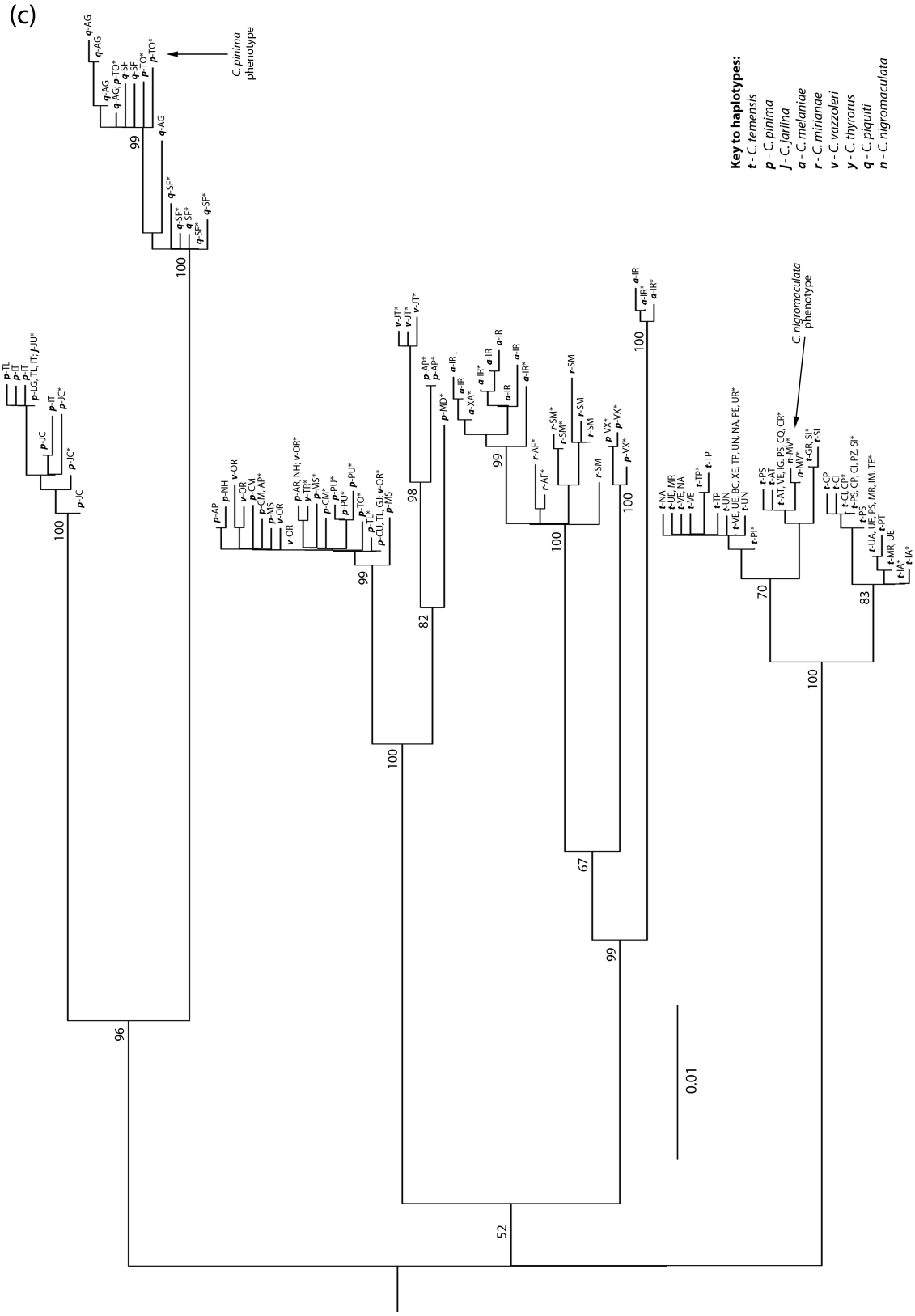


Figure 2 *continued*

(d)

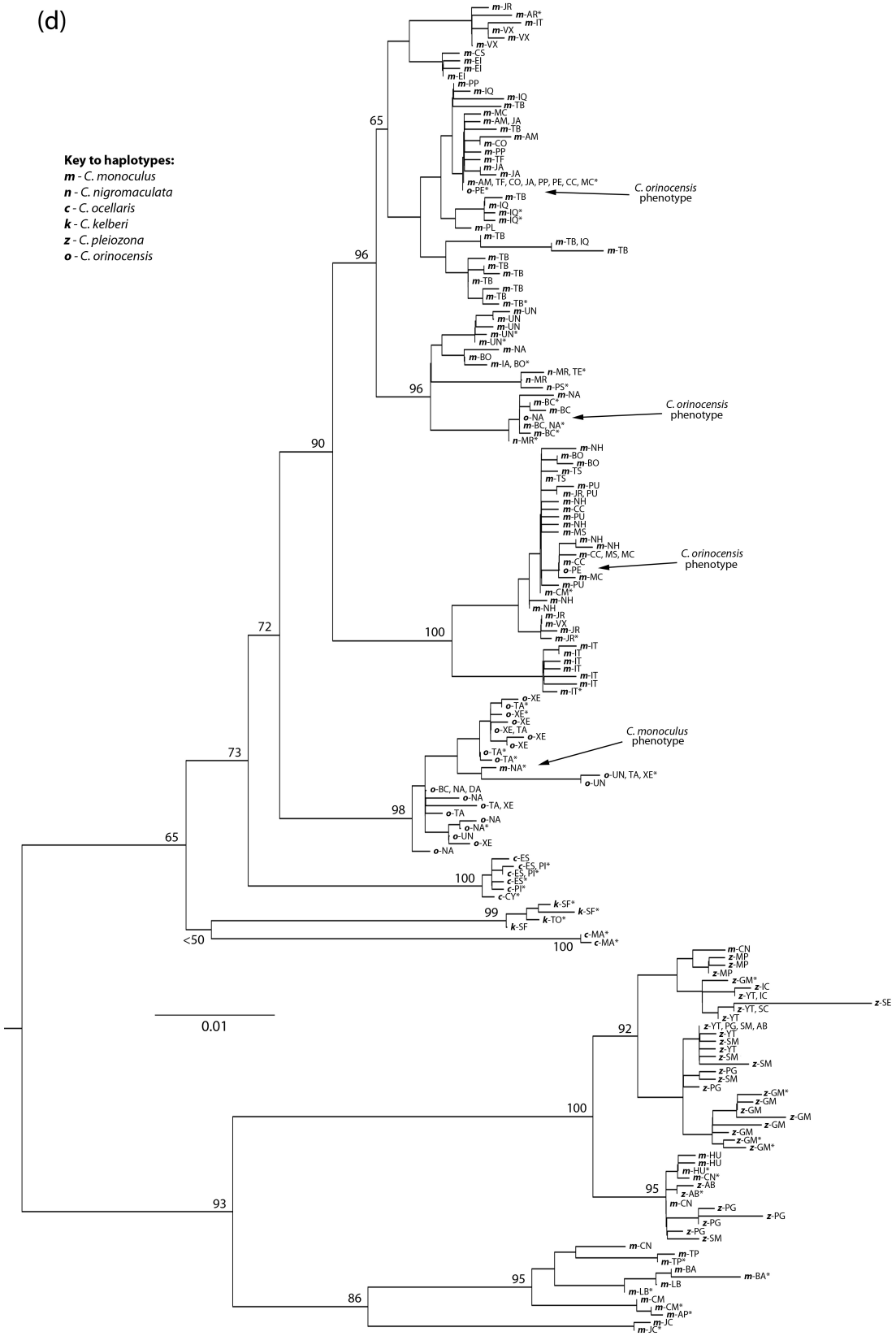


Figure 3

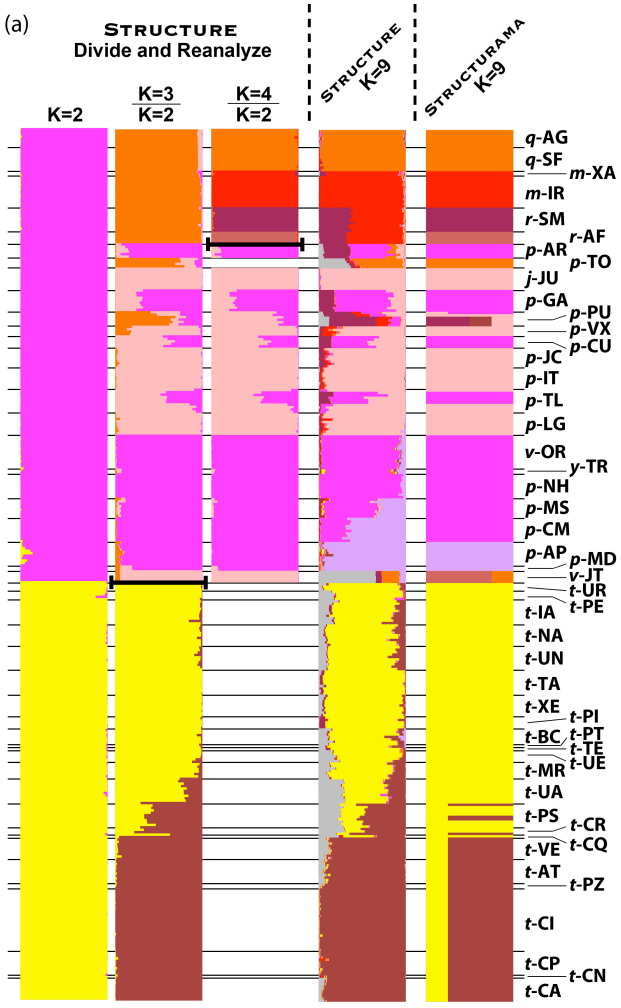

(a)
STRUCTURE
Divide and Reanalyze
K=2
K=3
K=2
K=4
K=2
STRUCTURE
K=9
STRUCTURAMA
K=9
q-AG
q-SF
m-XA
m-IR
r-SM
r-AF
p-AR
p-TO
j-JU
p-GA
p-PU
p-VX
p-CU
p-JC
p-IT
p-TL
p-LG
v-OR
y-TR
p-NH
p-MS
p-CM
p-AP
p-MD
v-JT
t-UR
t-IA
t-PE
t-NA
t-UN
t-TA
t-XE
t-PI
t-BC
t-PT
t-TE
t-MR
t-UE
t-UA
t-PS
t-CR
t-VE
t-CQ
t-AT
t-PZ
t-CI
t-CP
t-CN
t-CA

Figure 3 *continued*

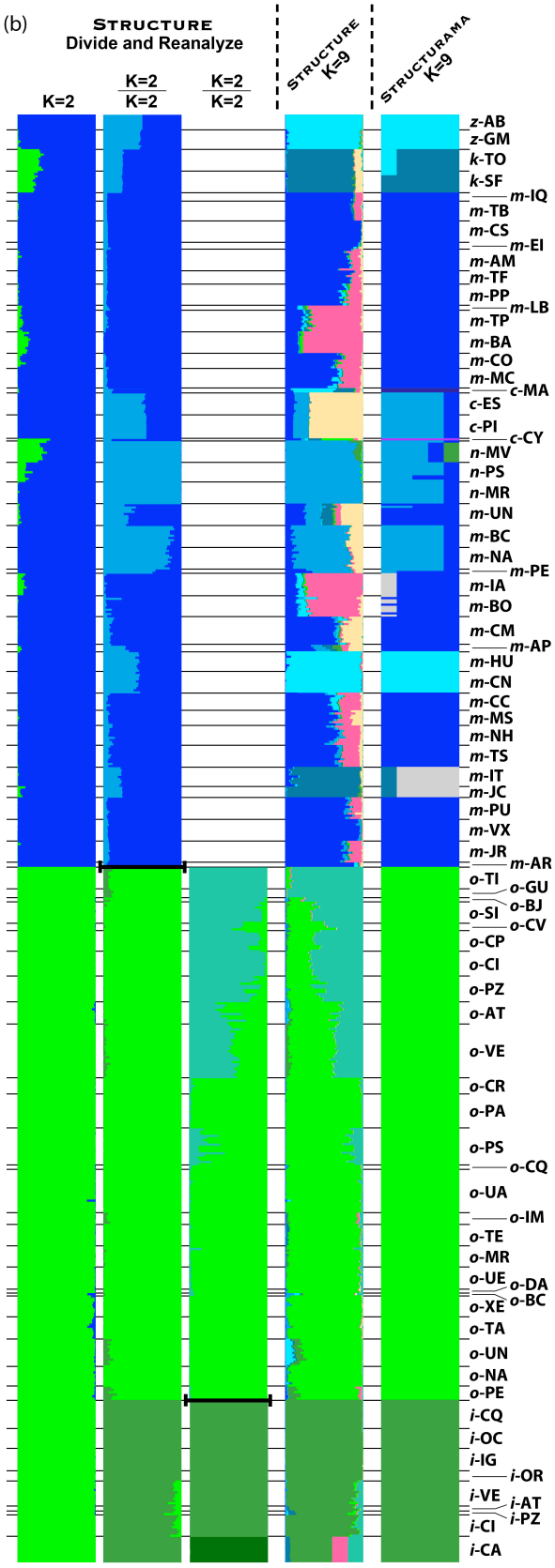


Figure 4

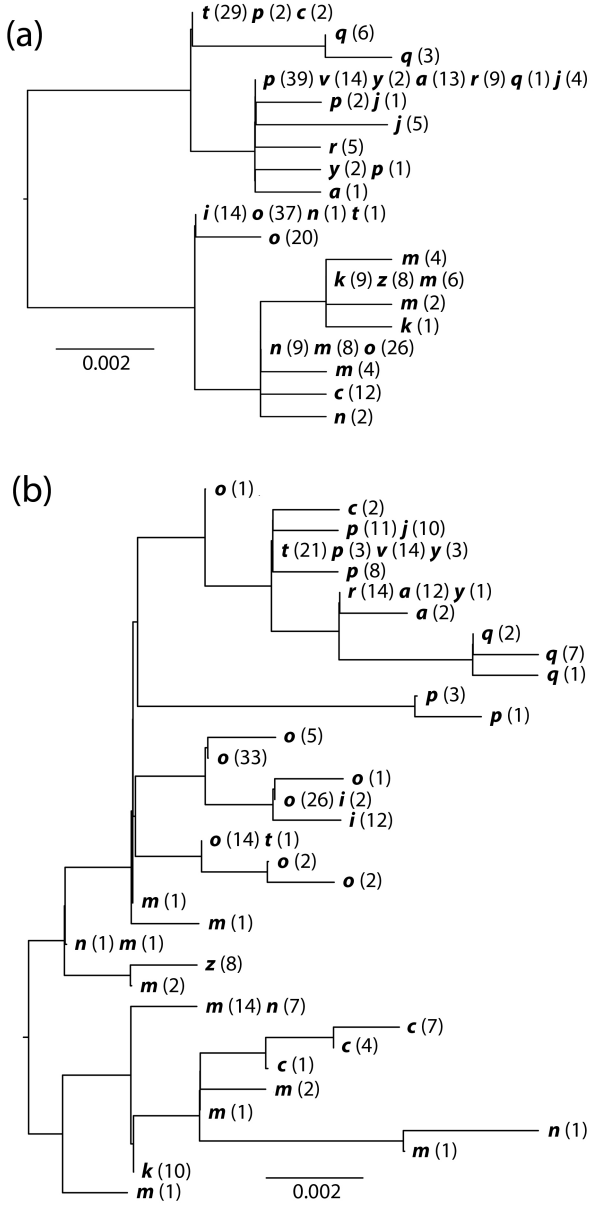
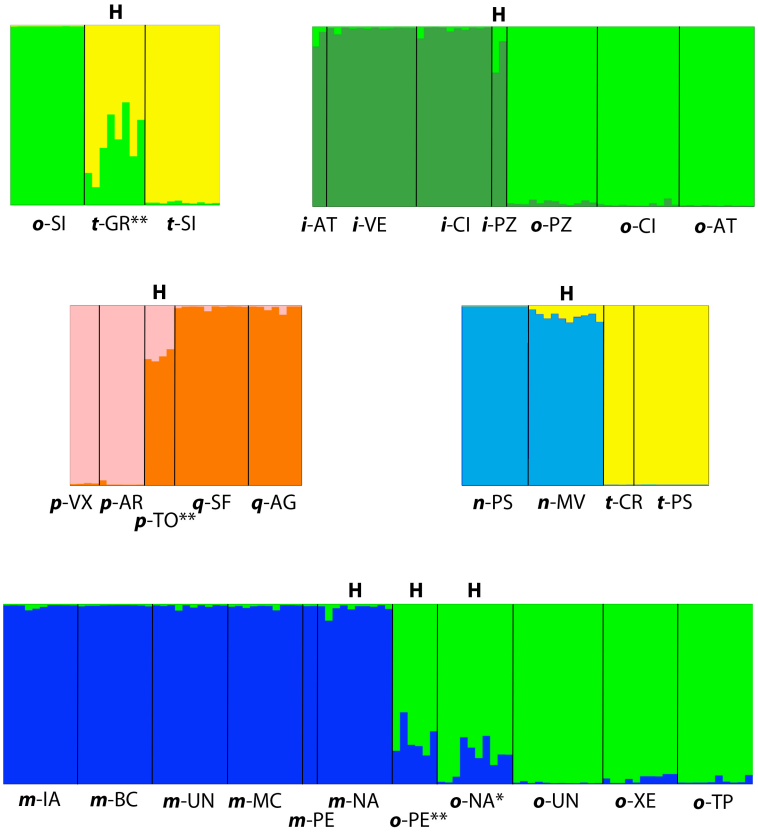

(a)
t (29) p (2) c (2)
q (6)
q (3)
p (39) v (14) y (2) a (13) r (9) q (1) j (4)
p (2) j (1)
j (5)
r (5)
y (2) p (1)
a (1)
i (14) o (37) n (1) t (1)
o (20)
m (4)
k (9) z (8) m (6)
m (2)
k (1)
n (9) m (8) o (26)
m (4)
c (12)
n (2)
0.002
(b)
o (1)
c (2)
p (11) j (10)
t (21) p (3) v (14) y (3)
p (8)
r (14) a (12) y (1)
a (2)
q (2)
q (7)
q (1)
p (3)
p (1)
o (5)
o (33)
o (1)
o (26) i (2)
i (12)
o (14) t (1)
o (2)
o (2)
m (1)
m (1)
n (1) m (1)
z (8)
m (2)
m (14) n (7)
c (7)
c (4)
c (1)
m (2)
m (1)
n (1)
m (1)
k (10)
m (1)
0.002

Figure 5


H
o-SI
t-GR**
t-SI
H
i-AT
i-VE
i-CI
i-PZ
o-PZ
o-CI
o-AT
H
p-VX
p-AR
p-TO**
q-SF
q-AG
H
n-PS
n-MV
t-CR
t-PS
H
H
H
m-IA
m-BC
m-UN
m-MC
m-PE
m-NA
o-PE**
o-NA*
o-UN
o-XE
o-TP

Chapter 3: Multilocus estimates of the species tree for the Amazonian peacock basses (Cichlidae: *Cichla*)

Authors: Stuart C. Willis, Izeni P. Farias, & Guillermo Ortí

Introduction

Comparative methods (Felsenstein 1985b; Harvey and Pagel 1991) provide a powerful means to test hypotheses of the history of evolutionary processes, particularly when applied to phylogenies where the tips are species (e.g. Collar et al. 2005; Whittall and Hodges 2007). However, species-level divergences are commonly subject to phenomena that complicate the inference of a robust phylogenetic estimate. Among these are the incomplete sorting of ancestral polymorphism or deep coalescence of gene lineages, porous species boundaries and transfer of hetero-specific gene lineages, and an insufficient degree of variation among homologous loci for the recovery of the tree (Maddison 1997). While these systematic biases can complicate phylogenetic estimation of long-diverged taxa, they are believed to be much more common among species-level phylogenies because of the recency of divergences involved (Edwards et al. 2007; Brumfield et al. 2008). Although typical phylogenetic approaches for such cases involve estimating phylogenetic history using loci that exhibit elevated rates of mutation and coalescence (typically mitochondrial DNA), the systematic literature has long emphasized the need to use multiple unlinked genetic loci due to the bias that any single marker may present regarding phylogeny (Maddison 1997). More recent approaches using this paradigm have demonstrated highly supported phylogenetic estimates based on the analysis of concatenated data from many separate loci (Rokas et al. 2003). However, it remains unclear just how many genes are necessary to infer an accurate and well-supported multilocus phylogeny (Rokas and Carroll 2005). In addition, it has long been recognized that each unlinked marker, those inherited separately via recombination, exhibits a separate and possibly incongruent gene tree (Pamilo and Nei 1988; Rosenberg and Nordborg 2002). Moreover, for some species trees with extremely short internal branches, the most common gene tree may not match the species tree topology (Degnan

and Salter 2005), and analyzing concatenated loci can result in inconsistent phylogenetic estimates (Kubatko and Degnan 2007). However, it remains to be seen how common species trees of this sort are, and in any case, recent work has suggested that rather than anomalous gene trees, the most common reconstruction for branches in the anomaly zone is a polytomy (Huang and Knowles 2009).

Neotropical freshwater fishes represent the most species-dense assemblage freshwater fishes, and perhaps 10% of all vertebrates species (Reis et al. 2003; Albert and Crampton 2005). However, the taxonomy and evolutionary biology of this assemblage is woefully under-studied, and no cohesive theory has yet emerged to explain the origins of this remarkable fauna (Hoorn et al. 2010). Among these colorful and charismatic assemblage of fishes, peacock bass of the genus *Cichla* are known for their voracious strike and powerful fight (Winemiller 2001). While a morphological review determined the presence of 15 described species in the genus (Kullander and Ferreira 2006), our recent study using extensive molecular data showed that several of these are better considered evolutionary significant units of more inclusive species (Chapter 2). We also observed that introgressive hybridization, while overall rare in terms of individuals, was widespread throughout the genus. Although a rough cytochrome *b* clock would place the earliest divergences among extant species in the late Miocene ~7 mya (Willis et al. 2010), time-calibrated analyses place these somewhat deeper in time (López-Fernández, in preparation) due to a relatively slow rate of molecular evolution in this lineage (López-Fernández et al. 2005). This may relate to the constraints imposed by a highly piscivorous lifestyle that are implied by the high degree of conservation seen in *Cichla* morphology (sensu Collar et al. 2009), and attempts at phylogenetic analysis of this genus using morphological characters have been largely ambiguous (Kullander and Ferreira 2006). To provide a basis for further study, here we estimate a molecular phylogeny for *Cichla* using nuclear loci.

Our phylogenetic reconstruction took two basic paths. First, we estimated a combined phylogeny using concatenated sequences from multiple loci. While it has been observed that more data (i.e. more genes) leads to a stronger phylogenetic estimate (Rokas and Carroll 2005), not accounting for the heterogeneous evolutionary rate or pattern among sites in the alignment can lead to systematic errors and inconsistent

phylogenetic estimates (Buckley et al. 2001; Bull et al. 2006). We attempted to accommodate potential variation in two ways. We allowed each gene in our alignment to reside in a separate partition, and each partition was allowed to have its own mutation rate and parameters. Because sites in each locus are physically linked and probably experience a similar cellular (e.g. proof-reading) environment, this partitioning scheme could accommodate much variance among sites. However, two problems plague this approach. With any *a priori* partitioning scheme, there is the risk of over-partitioning, and, therefore, over-parameterizing the data (Sullivan and Joyce 2005). This phenomenon occurs when there is insufficient information in a partition to accurately estimate the parameters of the model. The results are highly variable and inaccurate estimates of model parameters, including topology. In addition, these *a priori* partitions may not capture the largest variation in mutation pattern/rate variation among sites. For, example, while sites in different protein-coding genes are physically linked and inherited separately, the selective constraints and mutational biases often imply that sites in the first, second, and third codon positions among genes evolve in more similar ways than sites within genes (Reed and Sperling 1999; Willis et al. 2010). However, as our loci were largely non-coding, this potential partitioning scheme did not apply. An alternative to partitioning the data is to apply a phylogenetic mixture model (Pagel and Meade 2004). Rather than partition the data *a priori*, mixture models estimate the posterior weight that a site evolves under a number of separate models while simultaneously estimating the parameters of those models. As such, mixture models use the data itself to estimate similarity in mutation parameters among sites. However, the weights for each model are applied uniformly across all sites regardless of individual fit. Moreover, unlike with structure-based partitioning schemes, there is no *a priori* way to determine how many models to apply to a given dataset (Li et al. 2008; Roberts et al. 2009).

In addition to our concatenated analyses, we also implemented a Bayesian concordance analysis (BCA) (Ane et al. 2007). In this analysis, the gene tree for each locus is estimated separately, without any influence from other gene trees. A species tree, or concordance tree, is then built by taking into account the individual topologies from the genes. However, unlike other consensus methods such as gene tree parsimony (Slowinski and Page 1999; Page and Cotton 2000), a BCA takes into account the

uncertainty in each gene topology by summarizing the posterior distribution of topologies sampled in each gene tree analysis (Ane et al. 2007). In addition, for each branch in the phylogeny, BCA also determines what sample of input trees supports that branch, and predicts what proportion of the overall genome will likely support it based on the sampled trees. This method has the advantage that data-rich partitions with incongruent gene trees cannot dominate the analysis as in concatenated estimates. However, it remains unclear how BCAs perform in the inference of phylogenies at the species level, where not only are gene trees likely to be incongruent, but incongruence could be either masked or artifactually inflated (i.e. low signal to noise ratio) by the low levels of character variation among recently-diverged species.

Methods

DNA amplification and sequence assembly

We obtained tissues of all described species of *Cichla*. Tissue collection and DNA extraction were described previously (Chapter 2). For two to seven representatives of each species, we amplified 21 putatively single-copy and unlinked nuclear loci (Table 1). These loci were derived from the literature or developed specifically for this study, and included protein-coding exons and introns, microsatellite-flanking regions, and anonymous loci (Zardoya et al. 1996; Sides and Lydeard 1999; Won et al. 2006; Li et al. 2010). For those loci developed here, we obtained clone sequences from a library of genomic DNA enriched for microsatellite repeat motifs (see Macrander, Willis et al. in prep.). From those clone sequences that did not contain microsatellites, or which had a large region flanking the microsatellite, we searched for similar sequences on the NCBI nucleotide database using the BLAST algorithm (ncbi.nlm.nih.gov). To ensure that selected loci had a low degree of conservation and were unlikely to derive from an expressed or multi-copy gene, we excluded those loci that found significant matches from the BLAST search, also excluding those loci that had low but significant overlap with other cichlid taxa. For each locus, we designed PCR primers to amplify a section of DNA between 300 and 1000 base pairs. Using effective primer pairs, we amplified 2, 7, or all 38 individuals for these loci in turn, excluding those loci that exhibited properties of multi-copy genes (e.g. sites that were heterozygous in all individuals) or many null (non-

amplifying) alleles. Primers for loci developed here, and those described by Li et al. (Li et al. 2010) are listed as Supplemental Table 1 (Appendix).

Amplification for Mitf and Xsrc were described before (Chapter 2). For all other loci, PCR reactions contained 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 2.0 mM $MgCl_2$, 200 µM each dNTP, 0.1 µM each primer, 1.5 µL of 20 mg/mL bovine serum albumin (Fermentas), 0.5 U of Takara ExTaq polymerase (proof-reading exonuclease activity), and 3 to 4 µL DNA extract (10–50 ng/µL) in 30 µL reaction volumes. PCR primers are listed in Supplemental Table 1 (Appendix). Thermal cycling used a touchdown protocol of 1 min at 94°C, 35 cycles of 30 sec at 94°C, 30 sec at *X*°C, and 1.5 min at 72°C, followed by 10 min at 72°C., where *X* was 58°C for 15 cycles and 54°C for 15 cycles. Amplicons were sequenced in both directions at the University of Washington High Throughput Genomics Unit (http://htseq.org). Where necessary, such as in the case of heterozygous indel mutations, amplicons were cloned via bacterial transformation using the pGEM-T vector (Promega), and two to ten clones were sequenced to recover both alleles. Sequences for individual loci were assembled using CODONCODE ALIGNER (CodonCode Corp.). Sequences will be submitted to Genbank.

Three datasets were constructed. First, direct sequences for each locus were concatenated using MacClade 4.08 (Maddison and Maddison 1992), treating each locus as a character set. Second, matrices of individual loci were used directly in individual analyses (see below). Finally, alleles from individual loci were estimated using the non-recombination model of PHASE (Stephens et al. 2001) and a phase probability of 0.6. We did not include cloned sequences of known phase to inform the analyses, but where appropriate, compared the posterior PHASE estimates with these known sequences. Where phase probability was less that 0.6, we left these sites as ambiguous (Garrick et al. 2010). These largely corresponded to heterozygous singleton sites.

Phylogenetic analysis

To estimate the species-level phylogeny of *Cichla*, we performed a Bayesian analysis of the concatenated loci using MRBAYES v3.1.2 (Ronquist and Huelsenbeck 2003). Models of molecular evolution for each character partition (locus) were chosen using the AICc in MRMODELTEST (Akaike 1974; Nylander 2004). Each partition was

allowed unlinked model parameters and mutation rates. Branch lengths used the default exponential prior of 10, and all other model parameters were left as default except for the transition/transversion ratio of TmoM27, which we found difficult to estimate effectively due to high variance in the posterior distribution. Thus, we applied a Beta prior of (6,3) following recommendations in the MRBAYES website (http://mrbayes.csit.fsu.edu/Help/prset.html), which approximated the observed ratio of transitions to transversions. We made two simultaneous runs in MRBAYES, each with 15 chains and a heating value of 0.1. Each analysis was run for 10 million generations, sampling every 1000 generations, and discarded the first 5 million generations as burn-in. We ensured that parameters had reached a stable distribution using TRACER v1.4.1 (Rambaut and Drummond 2007) and the online version of AWTY (Nylander et al. 2008). To estimate the appropriate root of this tree, we took two approaches. First, we used data from another Neotropical cichlid fish, the putative sister-group of *Cichla*, *Retroculus* (López-Fernández et al. 2010), but were only able to obtain data for 5 of the 21 genes: Mitf, Xsrc, TmoM27, GnRH3i1, and Gpd2i1. This analysis was run as above. However, as this taxon is quite divergent from *Cichla*, we also performed analysis of the concatenated dataset using a molecular clock prior on branch lengths. It has been suggested that when there is not a severe rejection of the molecular clock assumption, this method provides an adequate estimation of the phylogenetic root (Huelsenbeck et al. 2002). We tested the appropriateness of the ultrametric assumption in two ways. First, we calculated the Bayes Factor comparing the two models using the concatenated data (Nylander et al. 2004). All harmonic means and Bayes Factors in this study were calculated using TRACER. Second, we examined the enforcement of a molecular clock on genealogies for individual locus using PAUP* v4b10 (Swofford 2002) and the optimal models and parameters estimated with MRMODELTEST. Trees used were those recovered from a constrained maximum likelihood search (see below). Comparisons were performed with direct sequences and phased alleles, and the significance of the differences was estimated using a likelihood ratio test of nested models (Huelsenbeck and Bull 1996).

In a concatenated analysis, it is important to understand how each gene contributes to the combined phylogenetic analysis, especially when there is significant

variation in informative sites among partitions as we observed here. One possibility is that the most variable loci dominate the analysis, while signal from the less variable loci is swamped. Alternatively, some loci may increase the artifactual signal such that the addition of loci actually decreases the phylogentic signal to noise ratio in the data, decreasing support for our recovered tree. We tested these conjectures in three ways. First, we made constrained searches of genealogies for the individual loci under maximum-likelihood using PAUP* and compared these to genealogies from unconstrained searches using the Shimodaira-Hasegawa test (Shimodaira and Hasegawa 1999). Constraints were designed to test the support in each tree for phylogenetic relationships among species. In this case, if sequences (or alleles) from two species always form a clade with respect to other alleles, it does not matter what topology terminals within this clade take, so long as the clade is recovered. Thus, we included several polytomies in our constraint tree that nevertheless reflected inferred species relationships (see Results). Constrained and unconstrained heuristic searches for each locus were conducted in PAUP* using mutation parameters estimated with MRMODELTEST. Searches consisted of 10,000 random-addition replicates, each limited to 10,000 branch-swapping iterations. Trees from each locus were compared using SH tests in PAUP* using 1000 RELL replicates. Searches were made for both direct sequences and alleles.

At the level of divergence observed in our loci, it is difficult to know how much of the strength in inferred individual genealogies results from real phylogenetic signal versus spurious mutational noise. A more pertinent question is how do these loci interact in a combined analysis. Traditional phylogenetic theory suggests that total evidence approaches facilitate the emergence of common phylogenetic signal while the influence of hypothetically random noise is diminished (Baker and DeSalle 1997; Wenzel and Siddall 1999). In order to understand this, we observed the changes in phylogenetic signal among reduced combinations of loci. Continuing from above, second, we created jackknife replicates of loci to examine the variation in phylogenetic signal for our unrooted tree. Using Microsoft EXCEL and PAUP*, we created 10 replicate datasets each of 5, 10, or 15 randomly chosen loci and concatenated these together. We analyzed these in MRBAYES runs of 5 million generations, sampling every 500 generations and treated

the first 2.5 million generations as burn-in. Each analysis used two simultaneous runs with 15 chains heated at 0.1. We then used filters in PAUP* to estimate the proportion of trees in the posterior distribution of each replicate that recovered clades from our 21-locus concatenated tree (i.e. posterior probability for our original branches). As we observed higher posterior probabilities for branches in our 21-locus concatenated tree with the molecular clock enforced, indicating steeper 'hills' in 'tree-space' with this model, we used this same prior here, expecting the resulting congruence/incongruence between analyses to be more discrete with this model. However, for the 5 gene analyses, we found that in order to keep the analysis in a part of tree space congruent with the molecular clock assumption, it was necessary to enforce a topological constraint for the first division of *Cichla*, clade A vs. clade B. We therefore excluded these branches from our topological filters. Finally, third, for comparison to these random locus groups, we created matrices with the 5, 10, and 15 most variable genes and performed similar MRBAYES runs and topological filters. However, there is the chance that conflict between most-variable, reduced-data matrices and the full, 21-locus matrix results not from reduced phylogenetic signal in the matrices with less data, but rather that the smaller matrices are recovering the real phylogeny while the 21-locus dataset is mislead by noise introduced from less-variable loci. In this scenario, one would expect phylogenetic signal from the most-variable, reduced-data matrices to be stronger (i.e. posterior tree space to be less diffuse) than for the full matrix. Thus, we also examined the posterior signal in each of these most-variable matrices for their own optimal trees. Specifically, we observed the individual posterior probability of the most common tree in the posterior distribution for each analysis. Further, we calculated the highest log clade credibility for each of the four matrices (three most variable plus all) using TREEANNOTATOR v1.5.3, part of the BEAST package (Drummond and Rambaut 2007). Both of these statistics give an estimate of how acute or diffuse 'tree-space' is for each dataset, or in other words, how well do the inferred phylogenies explain the data.

To estimate how our *a priori* partitioning scheme affected our estimate of phylogenetic relationships, we also implemented a phylogenetic mixture model using BAYESPHYLOGENIES (Pagel and Meade 2004). We made two runs each of two models using 4 or 10 general time-reversible (GTR) rate matrices and unequal base frequencies

plus gamma rate variation with four rate categories. These two model constructions bracket the number of parameters estimated for the MRBAYES partition model (49 and 121 vs. 87 respectively, not including branch lengths). We ran each model for 5 million generations, sampling every 500 generations. Convergence was estimated with AWTY and TRACER, and the first 3 million generations were discarded as burn-in.

We also constructed a species-level phylogeny using BUCKY v1.4, which implements a Bayesian concordance analysis (Ane et al. 2007). This analysis depends on examining the frequency of observing the same tree topology across genes, and the first step requires summarizing the posterior tree distribution of each locus, identifying the frequency of unique topologies, using MBSUM. First, we made two simultaneous runs in MRBAYES with all 38 taxa for each gene, using the models implemented in the concatenated analysis. Each analysis included 10 chains with a heating of 0.15 and was run for 30 million generations, sampling every 500 generations. As these analyses converged very early, we discarded only 10,000 of 50,000 trees from each run as burn-in, providing 100,000 trees per locus. In our initial tests using 38 taxa, we discovered that almost all topologies were unique within single-locus posterior tree samples. We determined that this result stemmed from the common sharing of alleles (sequences) across species and the fact that MRBAYES always samples fully resolved trees, meaning that a great number of random topologies were generated for those individuals that shared alleles. We worked around this issue by reducing the number of operational taxonomic units (OTUs) analyzed in our BCA. We reduced the number of taxa in individual-locus trees by choosing individuals as representatives of OTUs (delimited species from Chapter 2 and where possible major evolutionarily significant units) based on the unrooted concatenated tree (see Results). However, as we were concerned that our choice of representatives might affect the outcome, we created 10 replicate datasets in which one individual was randomly retained to represent an OTU. We used PAUP* to prune selected taxa from the posterior tree sets for each replicate, and summarized these 10 x 21 trees separately using MBSUM. For each of the 10 replicates, the 21 tree summaries were used as input for BUCKY, which was run for 1 million generations with 10 chains after 100,000 generations of burn-in. The α parameter, the prior prediction for the level of concordance among separate genealogies, was left as the default 1.0 for these analyses.

For 16 taxa and 21 loci, this prior produces a distribution of 1 to 8 topologies with a mode of 3 different underlying genealogies among the 21 loci. We calculated a majority rule consensus of the primary concordance tree of each replicate using PAUP* and calculated the mean proportion of loci supporting each branch in the consensus tree based on estimates across replicates.

Results

We were able to obtain a sequence for every individual for every locus except one: *Cichla pinima* from the Paru for locus 1835e6, for a matrix with only 0.12% missing data. Sequences from the 38 individuals showed a range of variability across loci (Table 1), but number of variable sites was overall relatively low for phylogenetic data, especially in comparison to mitochondrial data for these same individuals (Willis et al. 2010). Not surprisingly, models of evolution chosen for these loci were relatively simple, with the most complex being the HKY model plus invariant sites.

Bayesian analysis of the concatenated 21-locus dataset with MRBAYES provided a posterior distribution with a harmonic mean log likelihood of -20,286.9 (Figure 1). The rooting of this tree was supported by both the outgroup analysis (-21022.7) and molecular clock analyses (-20294.4) in MRBAYES. Comparison of the clock and unrooted (non-clock) analyses using Bayes Factors ($2B_{10} = 15$) provides "very strong" support for rejection of the molecular clock, following the updated table in Nylander et al. (2004). However, when individual loci were tested for clocklike nature using constrained genealogies in a likelihood ratio test, only one locus showed a significant deviation and then only for phased alleles (Table 1). We repeated our analyses without this locus (CteOI2) and obtained qualitatively similar results. Tree length for the posterior distribution is quite short overall (mean TL = 0.0399, with a 95% highest posterior density of $3.6\text{x}10^{-2} – 4.4\text{x}10^{-2}$), particularly compared to the trees including the outgroup *Retroculus* (mean TL = 0.13). Interestingly, the molecular clock tree was shorter than our unrooted tree (mean TL = 0.0315, with a 95% HPD of $2.8\text{x}10^{-2} – 3.5\text{x}10^{-2}$).

This nuclear phylogeny from the unrooted analysis shows much similarity to previous mtDNA genealogies (Chapter 2; see also Willis et al. 2010). It recovers the two main clades of *Cichla*, and all of the delimited species, and some sub-specific units (*C.*

pleiozona, *C. kelberi*, *C. ocellaris, C. nigromaculata*) are recovered as monophyletic. This is particularly noteworthy for those populations whose representatives came from distant localities, including: *C. temensis*, *C. intermedia*, *C. orinocensis*, *C. mirianae*, and *C. nigromaculata*. Importantly, the subspecies of *C. pinima sensu lato*, *C. pinima*, *C. thyrorus*, *C. vazzoleri*, and *C. jariina*, are monophyletic but still divided into two clades, not unlike the mtDNA tree or microsatellite clusters for these populations. Similarly, while the subspecies of *C. ocellaris sensu lato* are monophyletic, *C. ocellaris sensu stricto* and *C. nigromaculata* are nested within *C. monoculus*. However, in this nuclear tree, *C. kelberi* has taken the place as most-divergent subspecies rather than *C. pleiozona*. Also noteworthy is the recovery of *C. orinocensis* as monophyletic; even though it was observed to be a cohesive species using microsatellite data, it exhibits two polyphyletic clades of mtDNA, one of which is more closely-related to the *C. ocellaris s.l.* populations (Chapter 2; see also Willis et al. 2007). Trees from the outgroup and molecular clock analyses were very similar to the unrooted tree, except that the molecular clock tree recovered a different topology within *C. pinima s.l.* (Figure 1), and *C. orinocensis* was recovered as paraphyletic to *C. intermedia* in the outgroup tree (not shown). Further, support for most branches was slightly to moderately higher in the molecular clock analysis. Similarly, the most common tree in the posterior of the unrooted analysis (a.k.a. MAP tree) was encountered only 1.9% of the time, while for the clock model this was 10.3% of the posterior. This suggests that the optimal tree island is much smaller for the clock model than its unconstrained counterpart.

Topologies from individual loci constrained to match species relationships depicted by the concatenated tree, showed a range of incongruence compared to their unconstrained counterparts when compared with SH topology tests (Table 1). Constrained trees for several loci were significantly less likely at an alpha of 0.05, but only one locus was significantly incongruent for both direct sequences and alleles at an alpha of 0.01. If we apply a Bonferroni correction for multiple related tests, the operative p-value changes from 0.05 to 0.0023, implying that none of these trees is significantly deviant. Moreover, it is difficult to estimate the impact of noise to signal in these data-poor analyses. Therefore we examined the support for our concatenated tree using different subsets of data under a molecular clock. Not surprisingly, we observed that

posterior probability for branches in our 21-locus molecular clock tree (minus the branches for clade A vs. B) increased as the number of loci analyzed increased (Figure 2). Interestingly, the posterior probability from the 10 most variable loci was not significantly different from the 15 random loci (although this may have been conflated by a limited total number of loci). Nevertheless, even the 15 most variable loci did not exhibit the same level of posterior support as our 21-locus analyses, implying that the least variable 6 loci nevertheless added important phylogenetic signal to the dataset. More importantly, not only was each set of most variable loci less supportive of our 21-locus tree, they were also less supportive of their own optimal trees. For example, the MAP trees for the 5, 10, and 15 most variable loci analyses only made up 0.1%, 0.8%, and 1.8% of the posterior distribution, respectively (versus 10.8% for the 21-locus tree), and none of these topologies matched the maximum credibility tree from each posterior (MC tree, i.e. the single tree with the maximum product of posterior clade probabilities). For both the clock and unrooted analyses of the full dataset, the most common topology was also the MC tree. Similarly, the highest log clade credibility (the product of clade probabilities for the MC tree) was -8.59, -5.79, -4.95, and -2.37 for the 5, 10, and 15 most variable and full 21-locus analyses, respectively, showing that posterior support for branches was more diffuse in each of the reduced data analyses compared to the full dataset.

Our analysis of the 21-locus dataset using a mixture model using BAYESPHYLOGENIES returned improved log likelihood values over our partitioned analysis in MRBAYES for both the 4 and 10 rate matrix models (harmonic mean LnL -20,188.18 and -20,188.6, respectively). The tree lengths for posterior distributions of these models were comparable to the unrooted analysis ($3.03x10^{-2} – 4.68x10^{-2}$ and $3.30x10^{-2} – 4.76x10^{-2}$, respectively). Interestingly, the 10-matrix model provided no clear improvement over the 4-matrix model despite providing for more variance in mutation processes in the data. Although we observed that individual replicate runs converged to stable posterior tree distributions using AWTY, models and replicates appeared to converge on somewhat different trees. While these trees were largely concordant with our unrooted, partitioned tree, differences were observed in the topology of the *C. ocellaris sensu lato* group. For example, one replicate of the 4-matrix model placed *C. pleiozona*

sister to *C. monoculus* and *C. nigromaculata*, to the exclusion of *C. ocellaris sensu stricto*, while the other replicate recovered the partitioned topology, with *C. nigromaculata* plus *C. ocellaris s.s.* sister to the Igapo-Acu *C. monoculus*, but with *C. pleiozona* sister to the *C. monoculus* from Iquitos and the Araguari River (not shown). More importantly, although we observed convergence in the posterior trees, model parameters exhibited very low effective sample sizes (ESS) as calculated using TRACER, and the traces for tree length and model likelihood exhibited significant swings in value over the course of the runs. This suggests that these models were mixing very poorly despite using 9 heated chains in addition to the cold chain, and that these results should not be interpreted with confidence. Nevertheless, it is interesting to note that the greatest degree of ambiguity in our dataset centers on the topology within the *C. ocellaris sensu lato* species group.

The majority-rule consensus of the primary concordance trees from 10 replicates of the Bayesian concordance analyses showed general congruence with our unrooted, partitioned tree (Figure 3). Points of incongruence include the topology within the *C. ocellaris sensu lato* group, where *C. kelberi* no longer resides on the first branch to diverge. However, it is difficult to know if this incongruence stems from true incongruence in concatenated vs. concordance analyses or our choice of taxon representatives for terminals in the latter analysis. In contrast, we found that *C. pinima sensu lato* is still divided into two groups, with *C. pinima sensu stricto* polyphyletic between these, and that *C. piquiti* is the sister to this clade. In this tree, concordance values for support among loci are variable, but overall quite low, with the lowest support associated with branches that were observed least often in the primary concordance trees (Figure 3).

Discussion

Species-level phylogenetics

The inference of evolutionary relationships among closely related species is an important step in recovering the tree of life, but species-level phylogenies present special challenges. Among these are the gene tree-species tree conflicts that result from porous-species boundaries and the retention of ancestral polymorphism (Maddison 1997;

Maddison and Knowles 2006). Even when species do not exchange genes among them, short times between speciation events and large population sizes decrease the chance that ancestral variation will sort to fixation between successive speciation events (Pamilo and Nei 1988). The result is that gene trees may not match the underlying species tree, i.e. the branching patterns and divergence times between species, and for some species trees with short internal branches, the most common gene topology will not match the species tree (Degnan and Rosenberg 2006). Some studies have examined this phenomenon and suggested that in this part of parameter space, concatenation of sequences from multiple genes may be statistically inconsistent, that is, that more genes will tend to produce the wrong tree with greater support (Kubatko and Degnan 2007). The alternative to concatenation recommended in these circumstances is to infer species trees using procedures that evaluate gene trees with respect to predictions of a multispecies coalescent model (Liu and Pearl 2007; Liu et al. 2008; Kubatko et al. 2009). However, there has yet to be a formal test developed to determine if individual empirical datasets lie in the anomaly zone, and more recent simulation studies have suggested that given realistic mutation rates, the most common inference for trees in the anomaly zone is more likely to be a polytomy rather than an inconsistent topology (Huang and Knowles 2009).

In any event, it is important to understand how the phylogenetic signal from separate loci compare in separate and combined analyses (Bull et al. 1993). In combined analysis, we saw that although there was a degree of diminishing returns as loci were added to the analysis, these loci nevertheless contributed important phylogenetic signal to the dataset, even less-variable loci (Figure 2). Moreover, the most variable loci did not dominate the analysis, as none of these reduced datasets recovered the combined data tree. It remains to be seen if additional loci would change these results, but we observed that most nodes in the tree were very well supported, especially under a molecular clock prior. In separate analyses, we saw that concordance across loci was apparently quite low, although this result might be deceptive. When individual loci were compared to our 21-locus concatenated tree using constrained heuristic searches and topology tests, most loci showed some degree of incongruence (Table 1), although without further investigation it is difficult to know the nature of this incongruence. For instance, low signal to noise ratios in individual loci could cause incongruence in individual analysis but still allow the

emergence of a consistent phylogeny in combination (Wenzel and Siddall 1999). Moreover, when phylogenetic signal was compared while taking into account uncertainty in the topologies of individual loci, as in our Bayesian concordance analysis, a phylogeny very similar to our combined tree emerged. Although support for individual branches in this concordance tree were on average relatively low, we suspect that the low concordance values result, at least in part, from the relatively low number of variable sites across loci and the retention of ancestral alleles among species. As concordance is measured as the proportion of loci exhibiting a clade in their posterior distributions, it is important to note that, in addition to conflicting clades among gene trees, lack of resolution for a group of species in a gene tree (sharing alleles) should also result in low concordance values.

Another challenge for shallow phylogenies is the lack of nucleotide variation among species, potentially masking or artifactually creating conflict among gene trees. One strategy for dealing with this is to screen loci for variation and preferentially use those loci that exhibit more variation. However, according to our results, this strategy makes sense only insofar as it does not come at the cost of using more total loci. While we discovered that using more variable loci produced more posterior support than randomly chosen loci at each size of dataset, 10 random loci outperformed the 5 most variable loci, suggesting that in some cases more loci simply is better (Figure 2). It also remains to be seen if loci with a significantly elevated mutation rate produce a trustworthy signal of phylogeny. If the background variation among species' genomes is generally low even across genetic structural classes, as we observed here, it forces one to wonder why a particular locus would show a level of variation *higher* than most non-coding loci. Optimally, this pattern would simply represent a lower degree of conservation at that locus, but a number of other cryptic patterns such as unidentified paralogy, concerted evolution or diversifying selection, could also produce such elevated variation (Zhang and Hewitt 2003).

Another challenge related to low variability across loci is how to effectively analyze that variation. Studies have shown that under-partitioning data can have important effects in phylogenetic analysis, as variation among processes are forced to conform to fewer models (Brown and Lemmon 2007). The opposite phenomenon, over-

partitioning the data, can be equally risky, as analyses are unable to adequately estimate parameters for data-poor partitions and often provide results with a high variance and erroneous confidence (Rannala 2002; Sullivan and Joyce 2005; Brown and Lemmon 2007). In situations such as ours, a partitioned analysis with few variable characters per locus, the risk of over-parameterization could be high. Nevertheless, while we observed that mixture models, which stochastically apply a specified number of separate models to sites across loci, produced a substantial improvement in likelihood, our analysis with these models was insufficient to have confidence in these results. While our analyses could have been improved by increasing chain length, high autocorrelation among samples (i.e. low ESS values) implies that our chains were not mixing well despite a strong heating scheme, a problem that could also stem from inappropriateness of the model for our data. On the other hand, we observed high ESS values and low variance in parameter estimates for our partitioned model. In any event, we found that trees from all of our concatenated analyses were virtually the same, except for topology differences within some of our delimited species.

Phylogeny of Cichla

Our inferred phylogeny of *Cichla* shows much congruence with previous phylogenetic estimates, with some important differences. Perhaps most intriguingly, the topology of the clade A species was fully resolved except for the *C. pinima sensu lato* group, which is, nevertheless, monophyletic. In previous mtDNA genealogies (Willis et al. 2007; Willis et al. 2010), a number of basal polytomies and paraphyletic groups were observed in this clade (see also Chapter 2). The most striking of these, the two mtDNA clades of *C. pinima s.l.*, were still more or less observed in this tree, but they were observed to be monophyletic here rather than polyphyletic. In our analysis of species boundaries in *Cichla* using mtDNA, microsatellites and sequences of two nuclear genes, we inferred that described species in this group had shared gene flow too recently to be considered separate species. The complex geographic structure among localities was more suggestive of either one species with complex population structure, or two species with extensive hybridization. The pattern in this phylogeny is similarly complex. While two clades are recovered, each containing one or two of the other described species,

several of the *C. pinima sensu stricto* individuals have switched allegiances from their mtDNA clade (Chapter 2). We admit, of course, that the topology of this concatenated tree should not be interpreted too strictly, considering that the effects of gene flow (or hybridization) and recombination likely violate phylogenetic assumptions in this part of the tree. Nevertheless, this phylogenetic backbone provides a context from which more direct, coalescent-based tests of the 'one species vs. two species vs. four species' hypotheses can be tested (Knowles and Carstens 2007; Carstens and Dewey 2010).

Similarly, in the inferred phylogeny, *C. orinocensis* is recovered as monophyletic. While microsatellites portrayed this as a single, interbreeding species, individuals of *C. orinocensis* exhibited two exclusive but deeply-divergent mtDNA clades that were polyphyletic with respect to other species in clade B. If *C. orinocensis* is truly more closely related to *C. intermedia* than the *C. ocellaris sensu lato* group as portrayed in this nuclear phylogeny, a remaining question regards the origin of this polyphyletic mtDNA. While ancient mitochondrial capture seems a likely hypothesis (Willis et al. 2007), comparison with the distribution of coalescent patterns among other loci will be necessary to adequately test this hypothesis (Chapter 5). Also in clade B, we observed that the topology among *C. ocellaris s.s.*, *C. kelberi*, et al. showed a good deal of ambiguity among different analytical methods, and between the nuclear and mtDNA data. While it is difficult to know from the current results what portion of this ambiguity results from inefficient extraction or interpretation of phylogenetic signal, it is nonetheless congruent with out recent interpretation of these described species as evolutionary significant units of a more inclusive species (*C. ocellaris s.l.*) which exchange genes over evolutionary time (Chapter 2).

The phylogeny of *Cichla* recovered here using molecular data and Bayesian analyses shows less congruence with the parsimony-based phylogeny of Kullander and Ferreira (2006) based on their morphological dataset. Although based on only 11 characters and not well supported, one prominent result was the grouping of *C. intermedia* with species from our clade A. While *Cichla intermedia* does appear to show more elongate and streamlined body shape than other clade B species, this is likely a secondary adaptation for occupation of higher current velocity habitats (Jepsen et al. 1997; Winemiller 2001). Moreover, both mtDNA and scnDNA place *C. intermedia* in

clade B with high support (Figure 1). Thus, in this case it appears that the morphological characters examined by Kullander and Ferreira (2006) are subject to convergent evolution.

Conclusions

Despite having relatively few variable characters per locus and in total, we inferred a strongly supported phylogeny of *Cichla*. It appears that even when the overall mutation rate for a group is constrained, adding more loci can allow for the emergence of a combined phylogenetic signal. While there is some evidence from our results that preferentially utilizing nuclear loci that have a higher level of variability (i.e. lower degree of conservation or higher base mutation rate) can improve the support for a combined posterior tree, the phylogenetic signal from additional, less-variable loci should not be underestimated. We also observed that even though apparent conflict among loci produced incongruence with the combined phylogenetic estimate, analyses accounting for uncertainty in the topologies of individual genes, nevertheless, provided a phylogenetic estimate that was highly congruent with our concatenated tree. It remains to be seen which analytical method for concatenated data, partitioning or mixture-modeling, provides the most efficient means of extracting phylogenetic signal while accommodating heterogeneous mutational processes among separate loci.

Table 1. Loci examined in this study, and results from heuristic searches and likelihood ratio tests. [a] number of variable sites; [b] nucleotide diversity of phased alleles; [c] Shimodaira-Hasegawa test of constrained and unconstrained topologies for direct sequences and alleles; [d] likelihood ratio test of the molecular clock hypotheses on constrained topologies of direct sequences and alleles

marker	type	reference	base pairs	S[a]	π[b]	model	SH p-value[c] direct / allele	clock p-value[d] direct / allele
mitf	ORF+intron	Won et al. 2005	743	24	0.006	HKY	0.083 / 0.046	0.95 / 0.98
xsrc	ORF+intron	Sikes & Lydeard 1999	747	32	0.006	HKY+I	0.027 / 0.013	0.98 / 0.86
GnRH3i1	intron	Hassan et al. 2002	374	5	0.002	F81	0.078 / 0.057	0.99 / 0.99
Gpd2i1	intron	"	319	6	0.001	F81	1 / 1	0.99 / 0.99
1835e6	intron	Li et al. 2009	632	20	0.007	HKY+I	0.017 / 0.01	0.09 / 0.99
8680e2,3	intron	"	1099	40	0.009	K2P	0.048 / 0.054	0.96 / 0.99
14867e1	intron	"	870	30	0.006	HKY+I	0.029 / 0.074	0.99 / 0.99
18049e2	intron	"	407	12	0.006	F81	0.55 / 0.223	0.99 / 0.99
35564e5	intron	"	766	29	0.007	HKY	0.023 / 0.014	0.99 / 0.99
36298e1	intron	"	458	9	0.004	HKY	0.059 / 0.059	0.99 / 0.99
55305e1	intron	"	592	11	0.002	HKY	1 / 0.043	0.93 / 0.99
55378e1	intron	"	807	9	0.001	JC	0.062 / 0.062	0.86 / 0.99
TmoM27	μ-sat flanking	Zardoya et al. 1996	337	7	0.003	K2P	0.051 / 0.051	0.99 / 0.99
CorE7	μ-sat flanking	this study	516	9	0.005	HKY	0.253 / 0.028	0.99 / 0.99
CorF12	μ-sat flanking	"	380	15	0.006	F81	0.062 / 0.008	0.98 / 0.99
Cpiorcons	μ-sat flanking	"	378	7	0.003	F81	0.512 / 0.664	0.99 / 0.99
CinKA7	anonymous	"	797	9	0.002	F81+I	0.201 / 0.2	0.99 / 0.99
CmoME5	anonymous	"	494	19	0.006	HKY	0.085 / 0.047	0.73 / 0.77
CmoMJ3	anonymous	"	504	14	0.003	F81	1 / 0.059	0.74 / 0.99
CteOA7	anonymous	"	484	16	0.007	JC	0.244 / 0.112	0.99 / 0.84
CteOI2	anonymous	"	765	38	0.007	HKY+I	**0.004 / 0.003**	0.92 / **0.003**
ATPase 8,6	ORF (mtDNA)	Willis et al. 2010	842	153	-	GTR+I+Γ	-	-

Figure Legends

Figure 1. Phylogram of the majority-rule consensus from the posterior distribution of the unrooted (no outgroup, no molecular clock) Bayesian analysis of 21 nuclear loci. This tree differs from the MC/MAP tree only in that these show resolution among sequences of *C. orinocensis*. Values associated with nodes are posterior probability without / with the molecular clock prior. Asterisks denote nodes that were collapsed in the constraint tree for the individual locus searches. Numbers associated with terminals in the *C. pinima sensu lato* and *C. ocellaris sensu lato* species groups denote sub-specific taxon groups used in the taxon replicates for the Bayesian concordance analyses.

Figure 2. Graph of posterior support for the 21-locus concatenated molecular clock topology using reduced or full datasets. Bars are mean posterior support across nodes, and for random locus analyses, 95% confidence intervals across replicates are indicated.

Figure 3. Majority-rule consensus of the primary concordance trees from 10 taxon-replicate Bayesian concordance analyses. Values associated with branches are above branches: proportion of concordance trees with a clade; and below branches: the mean proportion of loci supporting a branch across replicates.

Figure 1

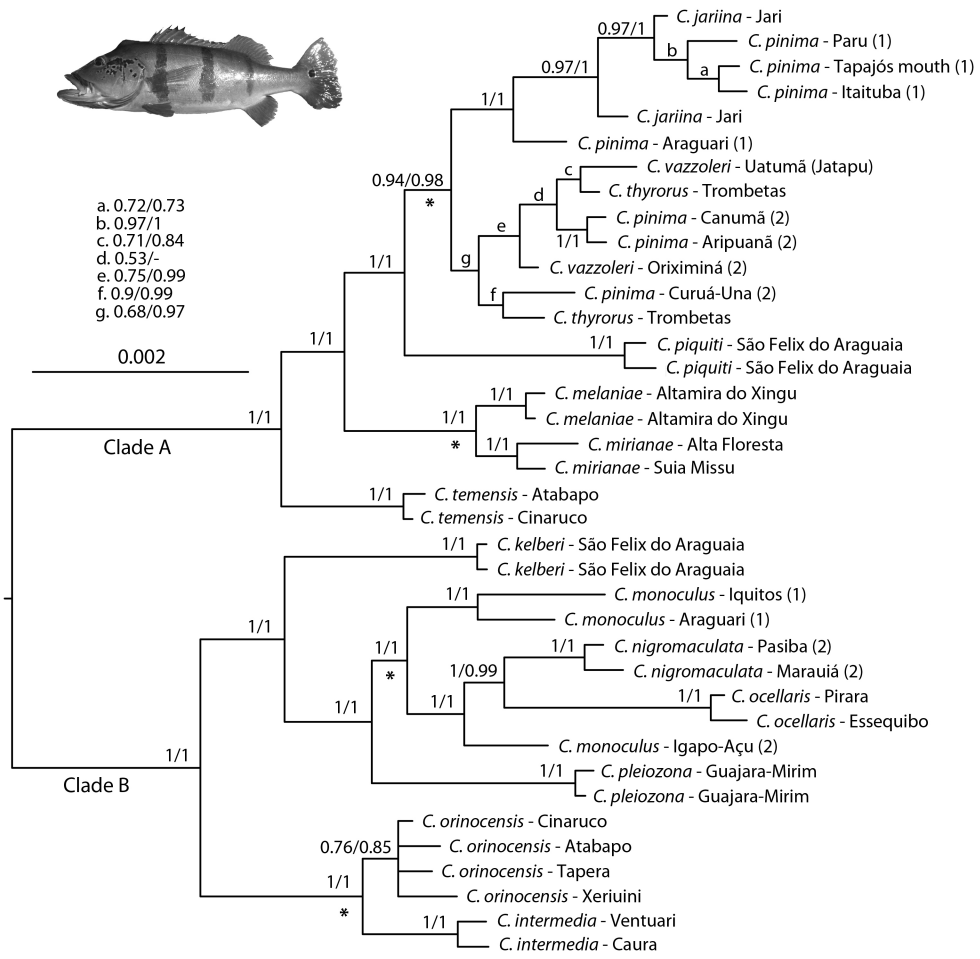

a. 0.72/0.73
b. 0.97/1
c. 0.71/0.84
d. 0.53/-
e. 0.75/0.99
f. 0.9/0.99
g. 0.68/0.97
0.002
Clade A
Clade B
C. jariina - Jari
C. pinima - Paru (1)
C. pinima - Tapajós mouth (1)
C. pinima - Itaituba (1)
C. jariina - Jari
C. pinima - Araguari (1)
C. vazzoleri - Uatumã (Jatapu)
C. thyrorus - Trombetas
C. pinima - Canumã (2)
C. pinima - Aripuanã (2)
C. vazzoleri - Oriximiná (2)
C. pinima - Curuá-Una (2)
C. thyrorus - Trombetas
C. piquiti - São Felix do Araguaia
C. piquiti - São Felix do Araguaia
C. melaniae - Altamira do Xingu
C. melaniae - Altamira do Xingu
C. mirianae - Alta Floresta
C. mirianae - Suia Missu
C. temensis - Atabapo
C. temensis - Cinaruco
C. kelberi - São Felix do Araguaia
C. kelberi - São Felix do Araguaia
C. monoculus - Iquitos (1)
C. monoculus - Araguari (1)
C. nigromaculata - Pasiba (2)
C. nigromaculata - Marauiá (2)
C. ocellaris - Pirara
C. ocellaris - Essequibo
C. monoculus - Igapo-Açu (2)
C. pleiozona - Guajara-Mirim
C. pleiozona - Guajara-Mirim
C. orinocensis - Cinaruco
C. orinocensis - Atabapo
C. orinocensis - Tapera
C. orinocensis - Xeriuini
C. intermedia - Ventuari
C. intermedia - Caura

Figure 2

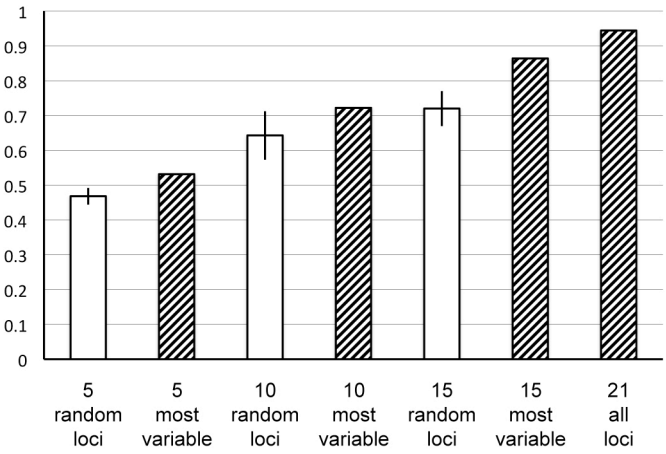
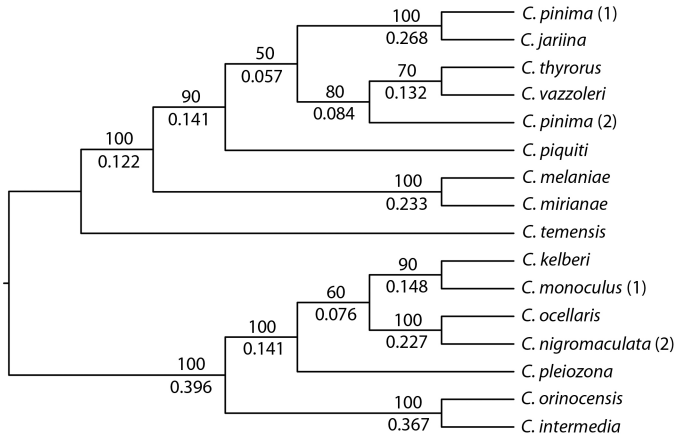


Figure 3



Chapter 4. The Casiquiare River acts as a corridor between the Amazonas and Orinoco River basins: Biogeographic analysis of the genus *Cichla*

Authors: Stuart C. Willis, Mario Nunes, Carmen G. Montaña, Izeni P. Farias, Guilllermo Ortí, & Nathan R. Lovejoy

Introduction

A fundamental task in evolutionary biology is determining the geographical context of speciation, and understanding how historical changes in geography have affected the divergence, dispersal, and extinction of lineages. The Casiquiare River of southern Venezuela is a unique biogeographic corridor between two of the most species-rich vertebrate faunas in the world (Reis et al. 2003; Winemiller et al. 2008). The Casiquiare provides a connection between the Amazonas, the world's largest river system, and the Orinoco, the river system draining most of Venezuela and Colombia (Rice 1921) (Figure 1). These two river systems contain extraordinarily diverse aquatic biotas and have long been considered separate biogeographic provinces and areas of endemism (Gery 1969; Weitzman and Weitzman 1982; Hubert and Renno 2006). Year-round, the Casiquiare captures flow from the headwaters of the Orinoco, but ultimately drains into the Negro River, the largest Amazonas tributary (Rice 1921; Lopez-Rojas et al. 1978). Thus, the Casiquiare appears to offer opportunities for both gene flow and dispersal between the aquatic biotas of the Amazonas and Orinoco (Mago-Leccia 1971; Freeman et al. 2007).

The Casiquiare region has a complex and dynamic paleogeographic history (Lundberg et al. 1998). Before ~10 Mya, a large "Paleo-Amazonas" River likely flowed northwards along the eastern edge of the Andes, entering the Caribbean in the vicinity of contemporary Lake Maracaibo (Hoorn 1994). This ancient river drained areas now occupied by the upper (western) Amazonas and upper and western Orinoco, which presumably composed a single biogeographic region (Figure 1). Indeed fossil fishes of many extant genera and species currently inhabiting the Amazonas, Orinoco, or both

have been found in geological formations from the "Paleo-Amazonas" period (Lundberg 1997, 1998). Between 10-11 Mya, uplift in the Eastern Cordillera of the Andes caused the Vaupes Arch to come into closer contact with the Andes, forcing the "Orinoco-Amazonas vicariance event" and separating the two drainages (Lundberg et al. 1998). Subsequent foreland sedimentation from Andean erosion forced the Orinoco to shift east where it took up its current position along the western edge of the Guyanan shield (where the current Casiquiare connection lies), while the Amazonas eventually broke through its eastern barrier, the Purus Arch, to take up its current path to the mid-Atlantic (Bermerguy and Sena Costa 1991; Hoorn 1994; Hoorn et al. 1995). These movements provided subsequent opportunity for drainage capture between the Orinoco and Negro headwaters to the east of the Vaupes Arch, and at some point the Casiquiare connection was formed (Figure 1). Unfortunately, the precise timing of the origin of the Casiquiare remains unknown (Stern 1970; Winemiller et al. 2008).

The Casiquiare region has likely played a role in both vicariance and dispersal of fishes between the Amazonas and Orinoco. The fragmentation of the "Paleo-Amazonas" and the associated Amazonas-Orinoco vicariance event likely isolated previously widespread taxa, resulting in allopatric sister lineages. More recently, the origination of the Casiquiare River may have subsequently allowed dispersal (range expansion) or gene flow between the Amazonas and Orinoco (see Winemiller and Willis 2010 for a discussion of alternative regional corridors). Current fish distributions reflect both these possibilities (Winemiller et al. 2008). Some putative sister species, such as the piranhas *Pygocentrus natteri* and *P. cariba* have allopatric distributions in, respectively, the Amazonas and Orinoco; these distributions could be explained by the Amazonas-Orinoco vicariance event. Other species, such as the piranha *Serrasalmus rhombeus*, the catfish *Phractocephalus hemiliopterus*, and the Amazonas River dolphin *Inia geoffrensis*, have distributions that include the Orinoco, Casiquiare, and Amazonas (Winemiller et al. 2008). These taxa may owe their broad distributions to dispersal through the Casiquiare. For example, Freeman et al. (2007) hypothesized that dispersal from the Orinoco to the Amazonas (via the Casiquiare) explains the distribution of the widespread piranha species *Serrasalmus manueli*.

For species with ranges that currently encompass both the Amazonas and Orinoco, a natural expectation is that the Casiquiare River could allow gene flow between the two basins. However, limited phylogeographic data for Neotropical fishes suggest that this is not necessarily the case. Lovejoy and de Araújo (2000) inferred that the Casiquiare does not permit gene flow between populations of the needlefish *Potamorrhaphis guianensis* distributed in the Amazonas and middle and lower Orinoco. Similarly, Turner et al. (2004) found populations of *Prochilodus rubrotaeniatus* from the Orinoco and Amazonas are genetically distinct, and showed no evidence that mitochondrial haplotypes are shared between these two basins. Possible barriers reducing the importance of the Casiquiare as a corridor have been suggested. These include the Atures rapids of the upper Orinoco (Chernoff et al. 1991; Lovejoy and de Araújo 2000), and an environmental gradient from neutral, unstained clearwater in the upper Orinoco to acidic, tannin-stained blackwater in the upper Negro (Sioli 1984; Winemiller et al. 2008). Such barriers might also explain why some fishes, such as the arowana (*Osteoglossum spp.*), arapaima (*Arapaima gigas*), and discus cichlids (*Symphysodon spp.*) are present in the Amazon, but are not present in the Orinoco.

Here we investigate the biogeographic role of the Casiquiare, including its impact on both gene flow and species distributions, using the cichlid genus *Cichla*. *Cichla* is a clade of 15 species of large (upwards of 12 Kg), colorful, and piscivorous fishes (Kullander and Ferreira 2006). Several species are economically important (Winemiller 2001), and are thought to play important roles in riverine food webs (Winemiller and Jepsen 1998; Layman and Winemiller 2004). *Cichla* is an ideal clade for examining the potential role of the Casiquiare because it includes species that are broadly distributed in all the drainages surrounding the Guyanan Shield. In addition, several species have distributions that span at least part of the Orinoco and Amazonas Basins and the Casiquiare. Previous work suggests that *Cichla* species' diversification has been shaped by the complex historical evolution of South America's hydrography from at least the Miocene (Willis et al. 2007).

We used intraspecific genetic data from populations of three *Cichla* species found in the Amazonas and Orinoco in combination with a species-level phylogeny to evaluate the biogeographic role of the Casiquiare. We sought to answer the following questions:

1) Does the current Casiquiare canal facilitate gene flow between contemporary populations in each basin, or if not, how long ago did populations diverge? and 2) What is the historical biogeographic origin of the current distributions of Orinoco and Amazonas species? To answer these questions, we first use mitochondrial sequence data from multiple populations of the three species with distributions that span the Amazon, Orinoco, and Casiquiare (*Cichla monoculus, C. temensis, and C. orinocensis*) to assess patterns of genetic structure and haplotype sharing, as well as the timing of divergence between populations (for distributions of these species see Figure 2a, 3a, and 4a). Second, for a deeper temporal perspective on the origin of biogeographic patterns, we reconstruct a phylogeny for the genus *Cichla*, and use this topology for dispersal-vicariance analysis (DIVA; Ronquist 1997). This analysis should illuminate the role of the Casiquiare in the dispersal of evolutionary lineages. By combining phylogeographic and phylogenetic approaches, we provide a synthetic assessment of the biogeographic role of the Casiquiare.

Methods

Intraspecific Analyses

For three focal species of *Cichla* (*C. temensis*, *C. orinocensis*, and *C. monoculus*), we obtained samples from localities throughout their distributions in the Amazonas, Orinoco, and Casiquiare (Figures 2a-4a; Supplemental Table 1). Voucher specimens were regularly taken (voucher data available upon request), but due to logistical constraints, many specimens were photographed, sampled nondestructively (dorsal fin), and released alive.

For each of the three species, we collected 500-550 bp from the 5' portion of the mitochondrial control region (CR), using previously described primers and conditions (Willis et al. 2007). Sequences from 76 new samples (Genbank GU295709 – GU295740) were combined with sequences generated for our previous study (DQ841819-DQ841946). In this previous study (Willis et al. 2007) we identified several haplotypes in *C. monoculus* (from the Mavaca River in the Orinoco basin) and *C. orinocensis* (from 3 localities in the Negro Basin: Unini, Xeriuini, and Tapera) that we inferred to be

introgressed from other species. These sequences were excluded from the following population level investigations.

For each of the three focal species, two separate alignments were generated using Clustal X (Thompson et al. 1997) with default parameters. The first alignment, used for phylogenetic (phylogeographic) analysis, used all unique CR sequences (haplotypes) for the focal species and one outgroup haplotype (selected based on Willis et al. 2007). The program Collapse (D. Posada-U. Vigo, Spain) was used to eliminate redundant sequences, with gaps treated as a fifth state and missing data ignored. The second alignment, used for population genetic (coalescent) analyses, included all sequences from the focal species, with no outgroup sequence. Data exploration suggested that modest variation in alignment parameters (gap opening/extension costs between 1 and 30) produced no change in recovered topologies, while higher alignment parameters produced biologically unrealistic alignments.

Phylogenetic analyses of haplotypes were conducted with Treefinder (Jobb et al. 2004) under maximum likelihood. In Treefinder, tree searches (1000 iterations) and bootstrap analysis (500 pseudoreplicates) were conducted under the model proposed by Treefinder. Treefinder was also used to execute SH tests of topologies constrained to fit different biogeographic scenarios (Shimodaira and Hasegawa 1999). We expected that reciprocally monophyletic groups of haplotypes in the Amazonas and Orinoco would indicate long-term isolation between populations in these two basins, while para- or polyphyletic haplotype lineages could suggest recent gene flow, or recent dispersal of a species from one basin to the other. Therefore, we tested for: 1) reciprocally monophyletic haplotype lineages in the Amazonas versus the Orinoco, 2) monophyly of Orinoco haplotypes and unconstrained Amazonas haplotypes, and 3) monophyly of Amazonas haplotypes with unconstrained Orinoco haplotypes. Each of these 3 constraint tests was performed with haplotypes found exclusively in the Casiquiare treated as members of either the Amazonas or Orinoco groups, for a total of 6 comparisons to the unconstrained tree.

While reciprocal monophyly of haplotypes in the Orinoco and Amazonas basins would suggest long-term isolation between these regions, para- or polyphyletic groups could indicate either recent gene flow or incomplete lineage sorting between isolated

populations. To discriminate between the latter two hypotheses, we used IMa, which implements a two-population isolation-with-migration model based on the coalescent (Hey and Nielsen 2007). The full IMa model has six parameters: population sizes for each of the two focal populations and the ancestral population, two migration parameters, and the time of divergence from a state of panmixia. In addition, using data from the full model, the program can examine nested models in which some of these parameters are equal (e.g. population sizes) or zero (e.g. migration) (we refer to these as "reduced" models hereafter). These models, when not significantly less likely than the full model, are generally preferred, to avoid over-fitting the data to a complex model and increasing variance in resulting parameter estimates. The "optimal reduced model" is the model with the fewest parameters that is not less likely than the full gene flow model. If reduced models with the migration parameters set to zero are not significantly less likely than the full model, we are unable to reject the null hypothesis of no gene flow between the Amazonas and Orinoco populations. For each species, data were analyzed using two models: (1) with migration (gene flow) permitted between populations and (2) with no migration (m=0). Since IMa only allows for the analysis of two populations, we conducted separate analyses for *C. temensis* and *C. orinocensis* with Casiquiare individuals grouped with either the Amazonas or Orinoco populations (referred to below as geographical groupings). These combinations meant that four different sets of analyses were run for each of these two species. In the examinations with gene flow (the "full" model, with six parameters), we also used the option in IMa to output the likelihoods of reduced models with fewer parameters. We compared the likelihoods of these to the full model using critical values from a X^2 distribution with appropriate degrees of freedom. In the results, we present the reduced models with the fewest number of parameters that were not significantly less likely than the full model. Unfortunately, it was not possible to examine reduced models in the no gene flow analyses, nor is it possible to directly compare these "forced" zero migration models with the full or reduced models because of the Bayesian MCMC construction of IMa. For the broadly distributed *C. monoculus*, we used only Negro River haplotypes to represent the Amazonas population in IMa. This provided better consistency with assumptions of the IMa model (panmixia within

populations), since samples from outside the Negro River represent distinct haplotype clades, indicating significant population substructure.

We analyzed our data under an HKY model in Bayesian runs constructed within the following ranges (depending on species): 100K to 200K genealogies sampled at an interval of 100 to 1000 generations with 20 to 30 chains under a geometric heating scheme, and following 3 to 10 million burn-in generations. For each analysis, numerous preliminary runs were made to determine the correct prior maxima and Bayesian conditions (heating scheme, length of burn-in, etc.) before two separate final runs were performed using different number seeds. From these final runs, we recorded the parameter estimates for the joint model. Analyses were constructed to ensure that effective sample sizes in the final runs were above 50, and where possible, posterior distributions were fully encompassed by prior maxima (see Results). To convert parameter values to meaningful biological estimates, we used a mutation rate for the CR calibrated in African cichlids (6% / Myr; derived from the divergence of two Lake Malawi cichlid clades (Sturmbauer et al. 2001)) and confirmed in Lake Victoria cichlids (Verheyen et al. 2003) and a generation time of five years (K. Winemiller pers. comm.).

Cichla *phylogeny and Dispersal Vicariance Analysis*

We evaluated *Cichla* phylogeny using three mitochondrial loci: CR, cytochome *b* (cyt *b*) and ATPase 8 and 6 (ATP). Portions of CR (500-550 bp) and cyt *b* (670 bp) were amplified and sequenced using primers and parameters described in Willis et al. (2007). Building on the dataset from Willis et al. (2007), we collected cyt *b* and CR sequence data for 102 new individuals, representing all remaining described species of *Cichla* (*C. mirianae*, *C. thyrorus*, *C.* cf. *vazzoleri*, *C. jariina*, *C. kelberi*, and *C. piquiti*), as well as additional localities for several species. New sequences have been deposited in Genbank as GU295666-GU295690 and GU295691-GU295708.

We also collected a new ATPase 8,6 dataset for *Cichla*. PCR primers for ATP were designed for this study: ATPLabF (5' AGCGTTAGCCTTTTAAGC 3') and ATPLabR (5' ACTATGTGGTATGCGTGTGC 3'). For amplification, 25 μL reaction volumes contained 20 mM Tris-HCl (pH 8.4), 50 mM KCl, 1.5 mM $MgCl_2$, 150 μM each dNTP, 0.4 μM each primer, 0.5 U of Invitrogen recombinant *Taq* polymerase, and 1 to 4

μL DNA extract (~10-50 ng/μL). Touchdown PCR was performed with an initial denaturation step at 94°C for 30 sec, followed by 30 cycles of 45 sec at 94°C, 60 sec at *X*°C, and 90 sec at 72°C, then a final extension of 5 min at 72°C, where *X* was 58°C for 5 cycles, 56°C for 5 cycles, 54°C for 5 cycles, and 52°C for 20 cycles on an MJ Research PTC200 thermal cycler. PCR products were sequenced at the University of Washington High Throughput Facility. New sequences have been deposited in Genbank as GU295741-GU295801.

The resulting dataset includes >2000 bp for 148 individuals representing all of the species recognized in Kullander and Ferreira's (2006) review of *Cichla,* although here we consider *C. nigromaculata* synonymous with *C. monoculus* (see Supplemental Figures for Chapter 1 and Supplemental Table 2 for sampling sites and species distributions). The cyt *b* and ATP sequences were trivial to align as no insertion-deletion events were detected. Individuals of each species with redundant sequences at cyt *b* and ATP were noted and eliminated from the matrix, leaving 73 unique haplotypes, and these were concatenated with their CR sequence. The CR haplotypes were aligned as described above. Concatenation of sequences from these regions is appropriate since mtDNA is inherited as a single, non-recombining unit and all sub-regions experience the same evolutionary history regardless of mutation pattern (Meyer 1993). The final matrix was submitted to TreeBase (SN4802).

To infer a genealogy of *Cichla* mitochondrial haplotypes, tree searches were implemented under parsimony (MP) and maximum likelihood (ML) criteria in PAUP* 4b10 (Swofford 2002) and Bayesian likelihood analysis (BA) in MrBayes v. 3.1.2 (Ronquist and Huelsenbeck 2003). Models for each partition were chosen using Modeltest 3.7 (Posada and Crandall 1998) or MrModeltest (Nylander 2004) for ML and BA respectively. Heuristic searches for ML included 1000 random addition replicates with unlimited tree bisection and reconnection branch swapping (TBR). This analysis was run with a single partition that included all data. Because MrBayes implements data partitioning more easily than PAUP*, BA was run with separate models of sequence evolution applied to each of 10 partitions corresponding to each gene (4, as ATP8 and 6 have different reading frames), and one each for 1^{st}, 2^{nd}, and 3^{rd} codon positions within the three protein-coding genes. For the Bayesian analysis, two simultaneous runs were

executed for 20 million generations with eight chains under default heating, sampling every 1000th generation, and the first 19,000 samples were discarded as burn-in. Convergence of runs in MrBayes was estimated using Tracer (http://evolve.zoo.ox.ac.uk/). MP analyses included a heuristic search with 10,000 random addition search replicates with TBR. Support for branches was estimated using 1000 bootstrap replicates (Felsenstein 1985a), each with 100 random-addition heuristic searches limited to 10000 rearrangements per random-addition replicate.

We inferred a species-level phylogeny by implementing the Minimize Deep Coalescences (MDC) method (Maddison and Knowles 2006), using the Deep Coalescences module in Mesquite 2.6 (Maddison and Maddison 2008). The Deep Coalescences module estimates the minimum number of gene lineages that must be postulated as deep coalescences (resulting from incomplete lineage sorting or minor ongoing gene flow) for a gene tree contained in a particular species-level topology, and allows Mesquite to commit heuristic searches of species trees that minimize the number of deep coalescences. As gene tree input, we used the maximum credibility tree from the BA (i.e. the tree with the greatest product of posterior probabilities), and instructed Mesquite to ignore branch lengths.

To generate probable historical biogeographic scenarios for *Cichla*, we used a dispersal-vicariance analysis (DIVA; Ronquist 1997). DIVA estimates the possible character states (distribution) of each ancestral node in the phylogeny by parsimony optimization, where divergence events between areas (vicariant speciation) and within areas (sympatric or microallopatric speciation) have a cost of zero, while dispersal or extinction events have a cost of one. Unlike other techniques to estimate ancestral distributions, DIVA does not force the distributions of ancestral species among geographical areas to be mutually exclusive, meaning that ancestral species can be distributed in more than one area at a time, as can contemporary species. As input, DIVA requires a fully-resolved species phylogeny and a distribution matrix of presence or absence of tip taxa in each geographic area. For the present analysis, species were scored as present or absent in areas defined by river basin: the Amazonas (including all tributary rivers), Orinoco, or Guyanas (including both the Essequibo and Maroni Rivers) (see Supplemental Figures for Chapter 1).

Results

Intraspecific Analyses

Cichla temensis

We obtained 546 aligned bp of CR from 128 individuals of *C. temensis* from 7 localities in the Amazonas basin, 9 localities in the Orinoco basin, and 1 locality in the Casiquiare (Figure 2a). These sequences collapsed to 22 haplotypes, none of which were shared between the Orinoco and Amazonas: 12 haplotypes were distributed exclusively in the Amazonas, and five were distributed exclusively in the Orinoco. Five haplotypes were found in the Casiquiare, of which one was exclusive to this location, three were shared with the Orinoco, and one was shared with the Amazonas. Treefinder proposed a TN+Γ model of evolution, and phylogenetic analysis resulted in a tree of likelihood -1139.329 (Figure 2b). In this phylogeny, haplotypes distributed in the Amazonas and Orinoco basins did not form reciprocally monophyletic groups, although haplotypes from the same or nearby localities were often clustered together. The haplotype unique to the Casiquiare was most closely related to a clade of haplotypes found in both the Orinoco and Casiquiare. When this exclusive haplotype was grouped with Orinoco haplotypes, SH topology tests showed that none of the three tested topologies (reciprocally monophyletic drainages, monophyletic Amazonas, monophyletic Orinoco) were significantly less likely than the geographically reticulate one we recovered by unconstrained heuristic search ($p=0.1966$, 0.2039, 0.1745, respectively). However, each of these constraints was less likely than the unconstrained topology when the Casiquiare-exclusive haplotype was constrained to group with Amazonas haplotypes ($p=0.0235$, 0.0227, 0.0216). Overall, the phylogeographic evidence suggests a history of genetic exchange between the Orinoco and Amazonas, presumably via the Casiquiare, as this locality exhibited haplotypes distributed across the tree and shared with both two major basins. However, no haplotypes were shared between basins outside of the Casiquaire River.

Our IMa analysis showed that reduced models that did not include gene flow were significantly less likely than the full model that did ($p < 0.05$) (Figure 2C). IMa indicated that a reduced model (with equal population sizes and equal and non-zero migration

rates) was not significantly less likely than the full model for both geographical groupings (2x log-likelihood ratio [2LLR] = 2.0486, 2.5505, 4.8532, 4.8623 with 3 d.f.). The alternative geographical grouping of individuals from the Casiquiare locality with either the Amazonas or Orinoco populations did not significantly alter the comparative likelihood of gene flow models. These results suggest that genetic exchange between the Amazonas and Orinoco is a better explanation for observed haplotype sharing than ancestral polymorphism. However, we note that we were unable to establish prior maxima that fully encompassed the posterior distribution of the time parameter in the full model for either geographical grouping; rather, the posterior distributions rose sharply to a plateau at approximately 400-500 Kya but then remained asymptotic over higher values. This is evident in the standard deviation associated with these values (Figure 2c) and suggests convergence to an equilibrium island model with open-ended divergence time (Supplemental Table 3); however, the minimum divergence time that is consistent with the estimated rate of gene flow between these populations is 400-500 Kya (see Discussion). Joint model analysis in IMa suggested that the divergence time between Amazonas and Orinoco populations was 1.45 to 1.75 Mya (with ongoing migration thereafter). Alternatively, without gene flow, divergence times were much younger (26 Kya-47 Kya), indicating that if we exclude gene flow from the model used to estimate gene genealogies, it is necessary to infer that populations were connected very recently. All parameters were fully bounded in the no-migration analyses.

Cichla monoculus

For *Cichla monoculus*, we obtained approximately 520 bp from 139 individuals from 16 localities in the Amazonas basin, 1 locality in the Orinoco basin, and 1 locality in the Casiquiare, reflecting the proportional distribution of this species in both basins (Figure 3a). The sequences collapsed to 50 haplotypes. Most of these haplotypes were found in the Amazonas basin; in fact, only a single *C. monoculus* haplotype was found in the Orinoco basin, and it was shared with all fishes from the Casiquiare locality. Treefinder suggested a TN+Γ model, and phylogenetic analysis produced a tree of likelihood -1515.904 (Figure 3b). In this tree, the haplotype from the Orinoco and Casiquiare was nested deeply within the Amazonas haplotypes, suggesting a recent

divergence or colonization. A topology where this haplotype was constrained to be sister to a clade containing the Amazonas haplotypes was significantly less likely than the recovered tree (SH test, p=0.0285). The asymmetric distribution of genetic diversity between basins, and the topology of the haplotypes, suggests a recent colonization of the Orinoco basin by *C. monoculus* from the Amazonas basin.

For IMa, we excluded analyses with the geographical grouping Orinoco+Casiquiare because this appeared to violate the IMa assumption of no reciprocal monophyly between populations. Analysis of the Casiquiare+Negro geographical grouping with IMa showed that a reduced model with no migration was significantly less likely than the full model, falsifying the hypothesis of no gene flow (Figure 3c). We found that two reduced models were not significantly less likely than the full model (2LLR = 3.2452 and 3.2040 for model (a), and 3.2459 and 3.2022 for model (b), both with 2 d.f.), but both of these models contained gene flow between the Orinoco and Amazonas populations. In fact, all models without bi-directional gene flow were less likely than the full model ($p < 0.05$). We were able to establish a maximum prior for divergence time in this analysis; this parameter had a unimodal distribution with a peak between 40 Kya and 75 Kya in both runs, but also had a wide right tail that increased the average. However we experienced difficulty finding an appropriate maximum prior for Amazonas (Negro) population size. The marginal distribution of this parameter showed a distinct unimodal peak that was fully encompassed, but the upper tail of the distribution reached a plateau at low but non-zero probability values for higher values of population size. Further, despite the peak in the marginal distribution, the joint estimate of divergence time between Amazonas (Negro) and Orinoco populations was 1.94 Mya and ~800 Kya in the full and reduced models respectively. In contrast, we found that when migration was set to 0, divergence time estimates were relatively recent (~34 Kya), albeit with a wide variance (Figure 3c). Very small effective population sizes were estimated for the Orinoco population which is not surprising given that a single control region haplotype was shared by all 17 fishes.

Cichla orinocensis

We sequenced 550 bp from a total of 105 individuals of *Cichla orinocensis* from 2 localities in the Amazonas basin, 9 localities in the Orinoco basin, and 1 locality in the Casiquiare (Figure 4a). The sequences collapsed to 26 haplotypes. Five of these were found only in the Amazonas basin, while 19 were only found in the Orinoco. The 10 Casiquiare fishes sequenced had 2 haplotypes: one of which was shared with the Amazonas (Negro), while the other was shared with the Orinoco. No shared haplotypes were observed between the Orinoco and Amazonas. Implementing a HKY+Γ model of evolution returned a tree of -1221.071 (Figure 4b). In this tree, Amazonas and Casiquiare haplotypes were nested among Orinoco haplotypes, but this topology was not significantly more likely than any of the constrained topologies ($p > 0.2$).

In IMa analysis of *C. orinocensis* data, models that did not include migration were significantly less likely than the full model ($p < 0.05$). The optimal reduced model had equal population sizes and migration rates, in both runs of both geographical groupings (2LLR = 3.3777, 3.2816, 5.2900, 5.3588 with 3 d.f.; Figure 4c). However, as in *C. temensis*, we were unable to find an effective maximum prior for divergence time in the full model. This parameter rose asymptotically to a peak between 650 Kya and 750 Kya for both geographical groupings and then stayed higher at increasing values of *t*. Consequently, divergence time estimates from the joint models (considering all parameters simultaneously across genealogies) with gene flow ranged from 1.78 to 1.91 Mya (with a wide variance), whereas divergence times without gene flow were much younger (~67 Kya) with low variance.

Species Phylogeny

After removal of redundant OTUs, our matrix for species phylogeny included 73 *Cichla* mitochondrial haplotype OTUs, each 2035 bp (670 cyt *b*, 158 ATP 8, 683 ATP 6, 524 CR). Maximum uncorrected sequence divergence among the *Cichla* haplotypes was approximately 7% for cyt *b*, 9% for ATP, and 14% for CR. In the alignment, 613 bp were variable and 434 were parsimony informative. Based on the relative rates estimated in the BA, CR had the highest estimated rate of molecular evolution, followed by 3rd, 1st, and 2nd codon positions across protein-coding genes.

Phylogeny estimation using MP, ML, and BA all recovered very similar trees (Figure 5), each with the same general topology as our previous analysis (Willis *et al.* 2007). ML recovered a tree with likelihood –9369.31448 under the GTR+I+Γ model chosen by Modeltest. BA found a model and tree with marginal likelihood –8757.94 using the following models for the 10 partitions chosen by MrModeltest (K80, JC, HKY, GTR, HKY, GTR+I, K80+I, F81, HKY+Γ, HKY+I+Γ). MP supported 2016 equally parsimonious trees, each with 1289 steps. These MP trees varied exclusively in the topology of haplotypes within species, with one exception. All three analyses indicate that *Cichla* species are partitioned into two main clades. The first of these (clade A) was comprised of *Cichla temensis*, *C*. cf. *vazzoleri, C. pinima*, *C. thyrorus, C. mirianae, C. melaniae*, *C. piquiti*, and *C. jariina*. Within clade A, some species did not exhibit monophyletic haplotype clades, although in only one case was a haplotype shared among species at all three loci (*C. pinima* and *C. jariina*). While the most common topology in the BA (PP 0.38) placed *C. jariina*, *C. piquiti*, and part of *C. pinima* as basal in Clade A, no analysis provided strong support for this arrangement: MP trees disagreed, and the optimal ML topology was a trichotomy. However, in no analysis was *C. temensis* at the base of the clade. The second clade of *Cichla* (clade B) contained two well-supported subclades: clade B1, including *C. monoculus*, *C. ocellaris*, *C. kelberi*, and *C. pleiozona*; and clade B2, including *C. orinocensis* and *C. intermedia*. The monophyly of clade B was well supported by MP bootstrap analysis, but only modestly supported by posterior probability. In addition, different methods produced minor rearrangements of closely related haplotypes, but these produced no effect on the subsequent biogeographic analyses.

Based on our mtDNA genealogy, a few species were not characterized by monophyletic clades of haplotypes (*C. pinima*, *C. mirianae, C. melaniae,* and *C. ocellaris*) that we interpret to result from incomplete lineage sorting (see Willis et al. (2007) for a discussion). To convert the genealogy to a fully resolved species-level phylogeny for use in DIVA, we used the Minimize Deep Coalescences module of Mesquite. MDC Analysis of the maximum clade credibility Bayesian tree produced 20 trees with 17 deep coalescences. We suspected that uncertainty regarding the placement of *C. pinima* was resulting in multiple topologies, so we removed this taxon and repeated

the analysis. This resulted in a single tree, to which we added *C. pinima* as indicated by the majority of its haplotypes (Figure 6a). As discussed below, our biogeographic conclusions are unaffected by this selection of the final tree or any of the 20 initially produced.

Dispersal-Vicariance Optimization

DIVA analysis of the species phylogeny resulted in 4 equally parsimonious optimizations, summarized in Figure 6. Each optimization requires 5 dispersal events and/or extinctions (these are equivalent in DIVA, much like accelerated and delayed transformation in parsimony optimizations). In each optimization, the inferred geographic distribution for clade A is identical: all internal nodes are reconstructed as Amazonas. For this clade, a single dispersal event into the Orinoco was inferred, explaining the current distribution of *C. temensis* (Amazonas and Orinoco). As mentioned above, this reconstruction is unaffected by alternative phylogenetic arrangements recovered in different species trees, as long as *C. temensis* is not placed as the basal species in clade A. In none of the 20 initial MDC species trees was this observed (but see Chapter 2).

In clade B, different possible histories were recovered for *C. monoculus*. Two optimizations (Figure 6a and c) indicate that this species dispersed from the Amazonas to the Orinoco after a vicariant divergence of its ancestor between the Amazonas and Guyanas. The scenario implied by the remaining two optimizations (Figure 6b and d) is that the common ancestor of *C. monoculus* and *C. ocellaris* was distributed in the all three biogeographic provinces (Amazonas, Orinoco, and Guyanas), and that vicariance separated *C. monoculus* (Amazonas, Orinoco) from *C. ocellaris* (Guyanas).

In the case of *C. orinocensis*, two optimizations (Figure 6a and b) indicate that *C. orinocensis* dispersed from the Orinoco to the Amazonas after a "non-vicariant" divergence of the common ancestor of *C. orinocensis* and *C. intermedia* in the Orinoco. Alternatively, *C. orinocensis* inherited its broad distribution from an ancestor that dispersed from the Amazonas to Orinoco (Figure 6c), or *C. orinocensis* dispersed directly from the Amazonas to the Orinoco (Figure 6d).

Interestingly, two different optimizations were determined for the common ancestor of clade B. Two optimizations (Figure 6a and b) suggested a history in which the

ancestor of clade B species dispersed from the Amazonas to the Orinoco, and the origin of clades B1 (*C. monoculus* et al.) and B2 (*C. orinocensis* & *C. intermedia*) resulted from vicariance between the Amazonas and Orinoco. The other two optimizations (Figure 6c and d) implied a scenario in which the common ancestor of clade B was distributed only in the Amazonas, with lineages from each sub-clade dispersing into the Orinoco independently.

Discussion

The Casiquiare River is a uniquely important biogeographic phenomenon because it provides a potential connection between two of the world's most species-rich vertebrate faunas (Reis et al. 2003). The origin of these faunas resulted not only from vicariance and isolation of lineages in the Amazonas and Orinoco basins, but also from exchange and interactions of lineages in secondary contact. Phylogeographic examination of genetic diversity provides a powerful tool for reconstructing species histories and the impacts of historical alterations of river drainage connectivity (e.g. Burridge et al. 2006, 2007). Here, mtDNA sequences from three species of cichlid fishes (*Cichla* spp.) provide strong evidence that the Casiquiare River has acted as a corridor for gene flow and dispersal between the Amazonas and Orinoco basins.

For the three species of *Cichla* examined, we discovered haplotype sharing across drainages, as well as poly- or paraphyletic distributions of haplotypes across drainages, implying relatively recent co-ancestry. For instance, in *C. monoculus* all fishes in the Orinoco basin possessed a single haplotype that was shared by all fishes in the Casiquiare. Further, this haplotype was closely related to haplotypes from the Negro and lower Madeira Rivers, and nested deeply within the remaining genetic structure of this species. This pattern suggests a recent dispersal of *C. monoculus* from the Negro into the Orinoco basin. In contrast, *C. temensis* exhibited complex genetic structure across basins, with some haplotypes more closely related to those in the opposing basin rather than those at nearby localities. This suggests a deeper history of population divergence with multiple instances of migration between basins. However, the relative importance of divergence and migration is difficult to ascertain from phylogenies alone.

To distinguish between ongoing gene flow and incomplete lineage sorting as explanations for the distribution of genetic diversity, we applied coalescent analyses as implemented in the IMa program. We found that optimal reduced models for each species included gene flow between populations in the Amazonas and Orinoco drainages. Although we had difficultly in estimating divergence time in several cases because the analyses began converging on equilibrium-island models, we found that when we did not consider gene flow divergence time was much younger. Joint parameter estimates ranged from 800 Kya to 1.91 Mya in the full and reduced models. In contrast, estimates from models without gene flow were less than 70 Kya in all cases (Figure 2c, 3c, and 4c). This difference in estimated divergence times between gene flow and no gene flow models follows from coalescent theory: if no gene flow is allowed between populations, any sharing of recently derived haplotypes must be explained by inheritance from the ancestral population (incomplete lineage sorting), a pattern promoted by recent divergence times and/or large effective population sizes. However, when gene flow is allowed, sharing of gene lineages can be analytically divided between inheritance and post-divergence genetic exchange, and divergence times become correspondingly older. In many cases though, and particularly where only one or a few loci are analyzed, it is possible for estimated gene flow to be so high that this masks the signature of inheritance from an ancestral population, and divergence times become effectively infinite (an equilibrium island model). In these cases, where divergence time is asymptotic, the mode of divergence time (unimodal distribution) or the approximate value at which the asymptote is first reached (when high divergence times have equal probability) can be informative. These values represent the minimum divergence times that are consistent with the estimated rates of gene flow. In the present study, these values were much lower than the average or joint model estimates (420-460 Kya for *C. temensis*, 680-730 Kya for *C. orinocensis*, and 40-75 Kya for *C. monoculus*), and much more similar to estimates made with the no gene flow models.

Our analysis of population structure using Isolation-with-Migration (IMa), has some shortcomings. The first is our use of a single genetic marker (mtDNA). It has been established that coalescent analyses are more accurate when based on multiple independent loci, and any single locus could provide a biased portrait of gene flow and

population structuring, such as in the case of sex-biased migration (e.g. Cooper et al. 2010). Ongoing studies with nuclear markers should reveal whether this is the case. Second, our analysis may have violated some of the assumptions of the isolation-with-migration model, principally the assumption of panmixia (no geographic structure) in each population. Population structure has the effect of preserving genetic diversity despite a reduced overall effective population size (sensu Wright 1943). In a coalescent analysis, this pooled genetic diversity may have the effect of inflating estimated effective population sizes (i.e. the Wahlund effect, Wahlund 1928; see also Lacy 1987; Harrison and Hastings 1996). It is unclear exactly what effect such inflated population sizes would have in an isolation-with-migration analysis, but recent studies have found IMa to be relatively robust to violation of the assumption of panmixia (Strasburg and Rieseberg 2008). These inflated population sizes may act to increase the probability of incomplete lineage sorting and decrease divergence time inferred under the current dataset, although the relative size of these effects may depend on the particular analysis and mutational distribution. Third, we were able to include relatively few localities in some portions of the range of each species. The Casiquiare, upper Orinoco, and upper Negro are remote and difficult to sample regions, limiting our ability to collect additional samples. However, our study is the first to examine species distributed across this region in a direct investigation of gene flow using parametric methods.

The shortcomings described above may have limited our ability to fully bound certain parameters with prior maxima or resulted in wide posterior distributions. Nevertheless, we believe that this does not invalidate important conclusions from our analysis. In the likeliest case, true rates of gene flow and divergence times lie between those estimated with each model variation (e.g. Peters et al. 2005; Hansson et al. 2008; Mäkinen and Merilä 2008). Thus, although we may not be able to rely on precise estimates of divergence time or quantitative rates of gene flow in either model, these results confirm that the Casiquiare region either serves as a conduit for contemporary gene flow between the Amazonas and Orinoco, or that gene flow across this region ceased very recently. Interestingly, the estimated rates of migration between basins are relatively low, on the order of a few alleles effectively exchanged per hundreds of thousands or million years (Fig 2c, 3c, and 4c). Based on the distribution of haplotypes in

the phylogeny and in space, it is clear that this level of gene flow has not prevented regions in each basin from developing their own endemic set of mitochondrial haplotypes (Fig 2b, 3b, and 4b). However, in most cases these haplotype groups are not reciprocally monophyletic by basin, and are separated by only a few mutations at a rapidly evolving locus, suggesting that gene flow was an important factor in the origin of this distribution.

Our intraspecific results indicate that the Casiquiare River has likely been responsible for gene flow between the Amazonas and Orinoco basins. This contrasts with previous studies that found no evidence for recent genetic transfer across this region (Lovejoy and de Araújo 2000; Turner et al. 2004). However, intraspecific analyses provide limited information regarding the formation of species distributions and biogeographic patterns over longer timescales. The question of whether lineages have dispersed between basins by using the Casiquiare and in what direction, as well as the possible effects of the Amazonas/Orinoco vicariance event, are best answered using our DIVA results.

Our recovered species phylogeny agrees well with our previous results (Willis et al. 2007), but includes species and localities that were not available for that analysis. Different analytical methods recovered very similar phylogenetic trees, indicating that the signal for most nodes is quite strong. The dispersal-vicariance analysis (Figure 6) provided unambiguous evidence that *C. temensis* dispersed from the Amazonas basin into the Orinoco. This finding is based on the exclusively Amazonas distribution of the clade A relatives of *C. temensis*. This scenario is reasonably consistent with the intraspecific data for this species (Figure 2), which show a polyphyletic distribution of mtDNA lineages between basins. However, the complex geographic structure of relationships between haplotypes across drainages suggests that this dispersal may have been followed by significant exchange of migrants.

In the case of *C. monoculus*, two different optimizations and presumed scenarios of dispersal and vicariance were reconstructed. The main difference between these scenarios is whether *C. monoculus* was ancestrally distributed in the Orinoco (i.e. already present there at the time it diverged from *C. ocellaris*), or more recently dispersed into the Orinoco. Evidence from the species' distribution, intraspecific mtDNA genealogy, and coalescent analysis clearly agrees with the latter inference (a young distribution in

the Orinoco, Figure 6a and c). First, *C. monoculus* has a relatively limited distribution in the Orinoco, as it is found only between Puerto Ayacucho and the Casiquiare. This limited distribution, despite wide habitat availability throughout the Orinoco (Winemiller 2001), is suggestive of a recent arrival in the basin. Second, if *C. monoculus* had been ancestrally distributed in both the Amazonas and Orinoco, barring demographic bottlenecks, we would expect to see much greater genetic diversity in the Orinoco basin. In contrast, only one of the 50 haplotypes observed in *C. monoculus* was found in the Orinoco basin, and this was shared with Amazonas fishes. As a result, it is not surprising that IMa estimated a very recent divergence between Orinoco and Amazonas (Negro) populations, using either the no migration or the migration model.

Finally, for *C. orinocensis*, DIVA was equivocal, providing three different possible scenarios: (1) dispersal from the Orinoco to Amazonas, (2) inheritance of an ancestral range that included both basins, or (3) dispersal from the Amazonas to Orinoco. Phylogeographic data provide some support for the former hypothesis. Haplotype diversity is higher in the Orinoco (although sampling effort may bias this pattern), and Orinoco haplotypes are relatively basal in the tree. Thus, we tentatively suggest that dispersal from the Orinoco to Amazonas is more likely than the reverse.

Considering the evidence from both intraspecific analyses and DIVA, we suggest that the best-supported biogeographical hypothesis for *Cichla* is depicted in Figure 6a. This scenario includes two dispersals of species from the Amazonas to the Orinoco, and a single dispersal from the Orinoco to the Amazonas. Importantly, in all cases, even the oldest IMa estimates of divergence time between Amazonas and Orinoco populations (<2 Mya) are considerably younger than estimates of the Amazonas-Orinoco vicariance event (~10 Mya, Lundberg et al. 1998). This further supports the idea that the distributions of our focal species are the product of dispersal events rather than inherited ranges corresponding to the ancient paleo-Amazonas. However, it is possible that the phylogeny of *Cichla* does exhibit some signal of the Amazonas-Orinoco vicariance. In Figure 6a, the ancestor to clade B is distributed in the Amazonas and Orinoco basins before undergoing allopatric speciation (vicariance). Fixing this node at 10 Mya implies a control region mutation rate of ~2.8%/Myr, much less than the published rates for cichlid fishes. Determining if the Amazonas-Orinoco vicariance event separated ancestral species in this

genus will require a time-calibrated multilocus phylogenetic analysis of *Cichla*. In any case, we suggest that our results support the conclusion that the combined use of intra- and interspecific data is a powerful approach for testing complex historical biogeographic hypotheses.

Although our current dataset, derived from three different species of *Cichla*, provides strikingly similar conclusions, the effect of any biogeographic feature depends on the ecological characteristics of the species impacted by it (e.g. Bermingham and Martin 1998; Thacker et al. 2007; Burridge et al. 2008). In a detailed study of the fish communities and environments of the Casiquiare, Winemiller et al. (2008) inferred that the Casiquiare River acts as a selective filter, where only some species are able to make the transition from stained, acidic water in the Negro River to unstained, neutral water in the upper Orinoco River. While *Cichla* species exhibit distinct habitat preferences (Jepsen et al. 1997), it appears that the species we studied are able to disperse across less suitable habitats and utilize the Casiquiare as a dispersal corridor. However, the three *Cichla* species with cross-drainage distributions do not show identical phylogeographic patterns or demographic estimates, implying that the timing and functional use of the Casiquiare River is mediated by the ecology of each species. Thus, to understand the utility of the Casiquiare region for the Amazonian ichthyofauna as a whole, it will be important to assess the importance of this dispersal corridor for taxa with varying ecophysiological requirements.

Figure Legends

Figure 1. Map of northern South America showing relevant geographic and geological features. Structural arches: Vaupes Arch; Purus Arch. Rivers: Orinoco, Essequibo, Maroni, Amazonas, Negro, Casiquiare. Major geological features: Guyanan Shield, Brazilian Shield, Andes Mountains. Tertiary River Drainage patterns: the "Paleo-Amazonas" River (based on Lundberg et al. 1998b) that historically connected the Amazonas and Orinoco Basin regions.

Figure 2. Intraspecific results for *Cichla temensis*. a) Map of northern South America showing approximate species range and sampling sites for the current study. Locality abbreviations follow Supplemental Table 1. b) Maximum likelihood phylogram of haplotypes with bootstrap support proportions greater than 50%. For each haplotype, the locality, number of times observed, and drainage occupied are shown. c) Results from IMa analysis. For each model and geographical grouping, parameters are as follows: log probability of the joint model (Log(p)), divergence time in millions of years (*t*), standard deviation of divergence time (σ of *t*) in millions of years, effective population size of Orinoco (θ_O) in individuals, migration rate of mtDNA from Orinoco to Amazonas ($m_{O \rightarrow A}$), effective population size of Amazonas (θ_A), migration rate of mtDNA from Amazonas to Orinoco ($m_{A \rightarrow O}$), and ancestral effective population size (θ_{OCA}). Models are: full model (six parameters), optimal reduced model, and model with migration forced to be zero. Values shown reflect a mutation rate of 6% sequence divergence per million years.

Figure 3. Intraspecific results for *Cichla monoculus*. As in Figure 2.

Figure 4. Intraspecific results for *Cichla orinocensis*. As in Figure 2.

Figure 5. Maximum credibility phylogram of *Cichla* haplotypes based on 2035 bp of the mitochondrial control region, ATPase 8,6 and cytochrome *b* loci. Values associated with branches are parsimony bootstrap and Bayesian posterior probabilities respectively. Branches marked with an asterisk had 99 or 100% for bootstrap values, and 1.0 for posterior probability.

Figure 6. Results of the Dispersal-Vicariance Analysis based on the *Cichla* species phylogeny inferred using Minimize Deep Coalescences. For each extant species, distribution in the Amazonas basin, Orinoco basin, and/or Guyanas drainages (e.g. Essequibo) is indicated, and focal species are underlined. Inferred distributions for ancestral species are shown at internal nodes. Dispersal, vicariance, and extinction events are shown on respective branches, and the order of events proceeds from bottom to top along each branch. Only the right portion of the phylogeny, Clade B, is shown for scenarios b through d. The optimization for Clade A is identical in all cases.

Figure 1

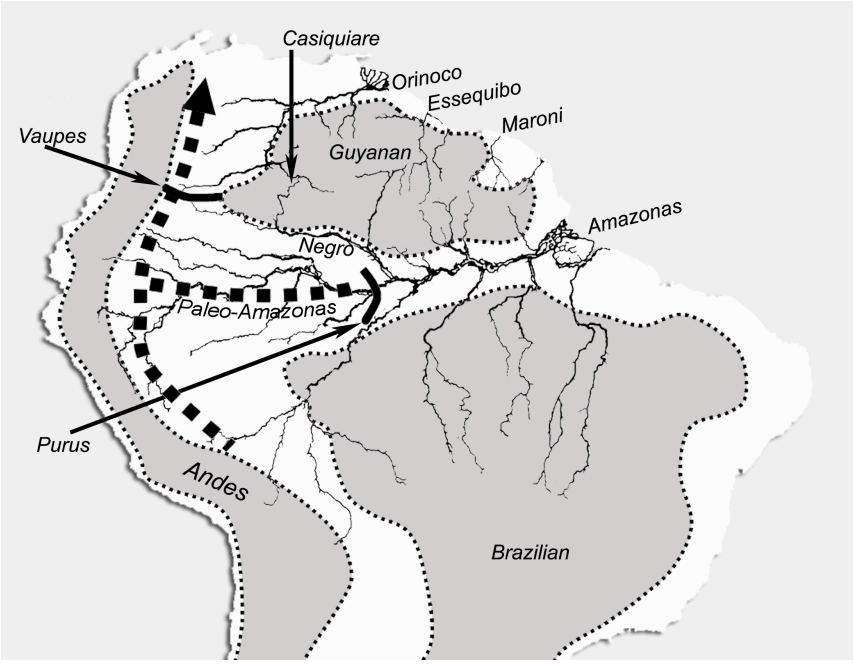
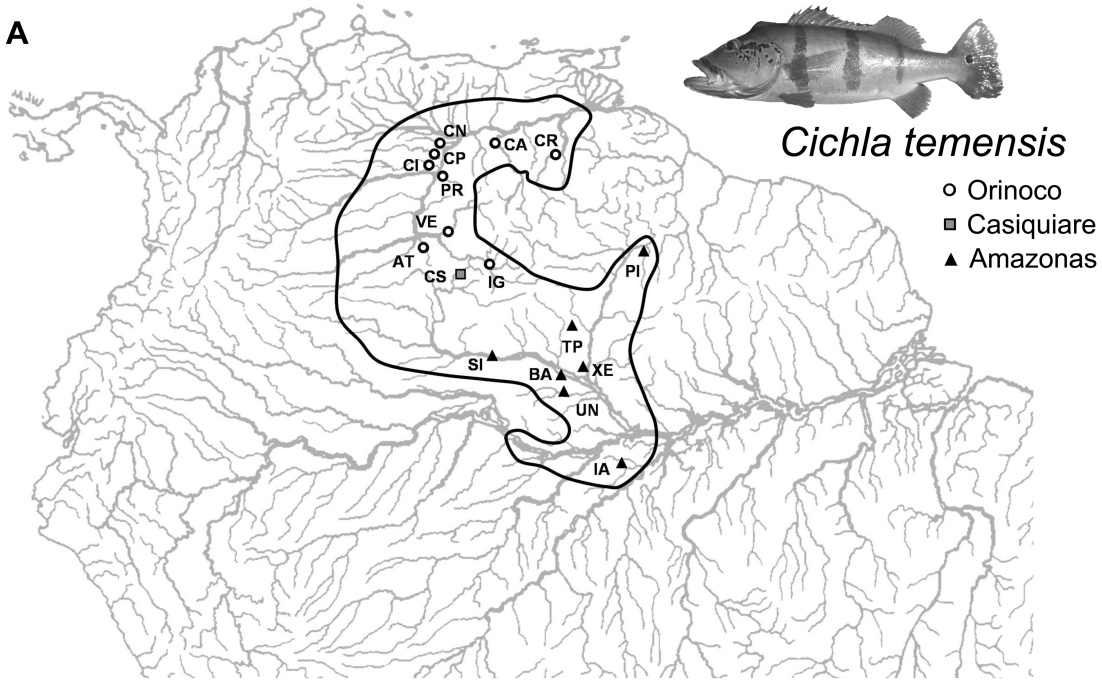
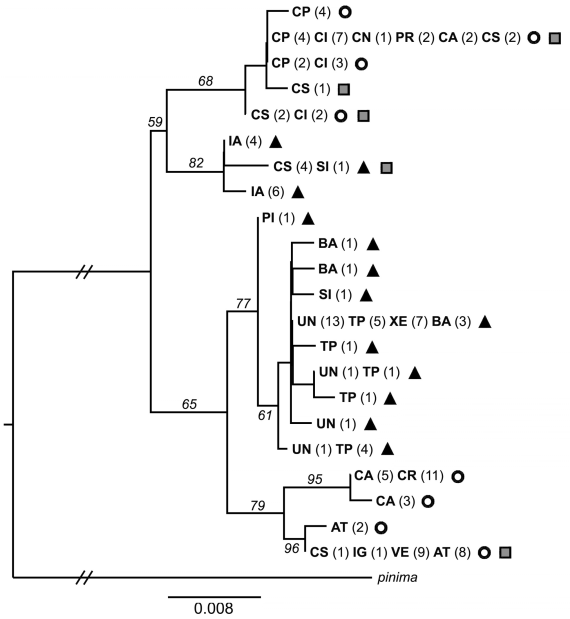

Casiquiare
Orinoco
Essequibo
Maroni
Vaupes
Guyanan
Amazonas
Negro
Paleo-Amazonas
Purus
Andes
Brazilian

Figure 2

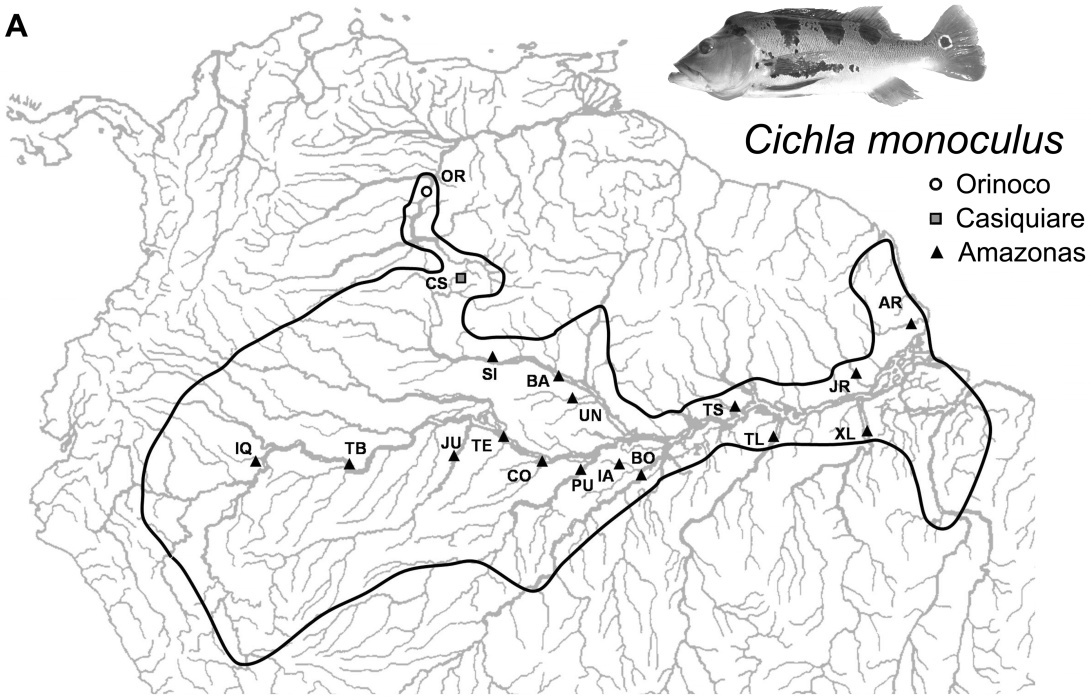


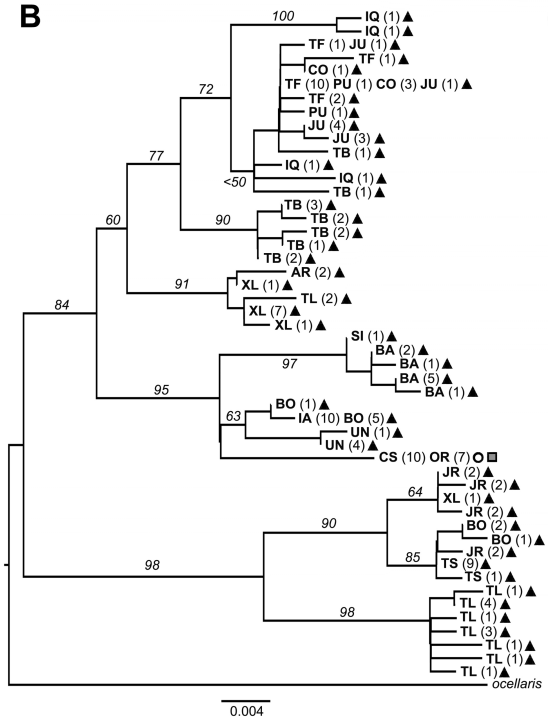


C

		model		
		full	reduced	m=0
Ori vs. Cas+Amaz	Log(p)	-9.2674 -8.2003	-11.6285 -10.6315	-8.6305 -8.6306
	t	1.62	1.75	0.026
	σ of t	1570	120	24
	θ_O	35400	39300	20400
	$m_{AC\rightarrow O}$	1.35×10^{-7}	3.85×10^{-6}	0
	θ_{AC}	51100	θ_O	76900
	$m_{O\rightarrow AC}$	6.44×10^{-6}	$m_{O\rightarrow AC}$	0
	θ_{OCA}	93	θ_O	86200
Ori+Cas vs. Amaz	Log(p)	-7.1003 -7.1254	-8.1246 -8.4007	-8.6490 -8.6488
	T	1.45	1.60	0.047
	σ of t	237000	13312	15
	θ_{OC}	45900	38100	32200
	$m_{A\rightarrow OC}$	3.64×10^{-6}	2.10×10^{-6}	0
	θ_A	51500	θ_{OC}	44200
	$m_{OC\rightarrow A}$	2.06×10^{-6}	$m_{OC\rightarrow A}$	0
	θ_{OCA}	28300	θ_{OC}	110500

Figure 3

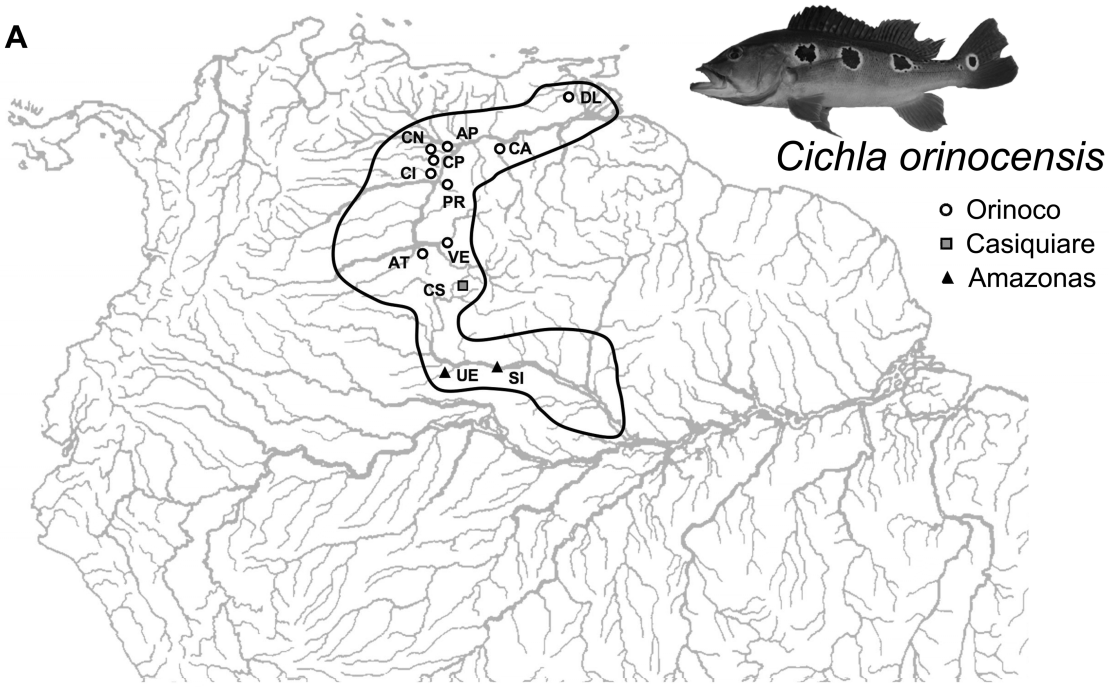


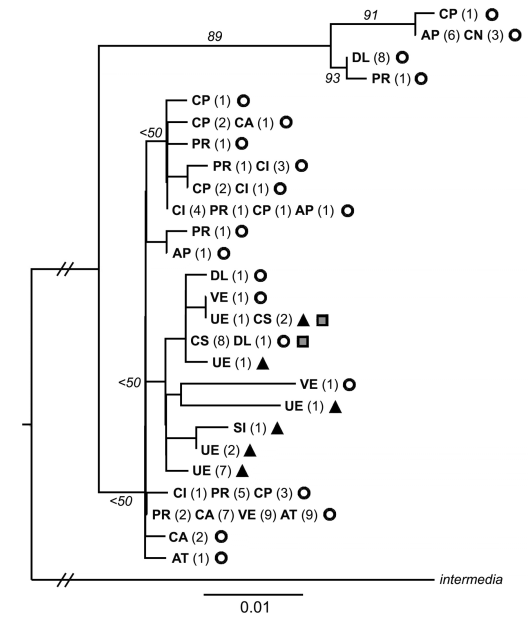


C

		model			
		full	reduced (a)	reduced (b)	*m*=0
Ori vs. Cas+Amaz	Log(p)	-12.4927 -14.1995	-14.1154 -15.8015	-14.1157 -15.8006	-9.6800 -9.6368
	t	1.94	0.81	0.802	0.034
	σ of *t*	88900	460000	454000	10300
	θ_O	60	420	420	20
	$m_{AC\rightarrow O}$	1.9×10^{-5}	2.33×10^{-6}	2.36×10^{-6}	0
	θ_{AC}	79700	27400	27400	28500
	$m_{O\rightarrow AC}$	5.51×10^{-6}	$m_{O\rightarrow AC}$	$m_{O\rightarrow AC}$	0
	θ_{OCA}	74789	θ_O	θ_{AC}	36800

Figure 4





C

		model		
		full	reduced	*m*=0
Ori vs. Cas+Amaz	Log(p)	-7.3672 -7.5411	-10.0122 -10.2205	-9.6438 -9.6576
	t	1.91	1.78	0.067
	σ of *t*	40500	27600	250
	θ_O	76100	67200	92800
	$m_{AC\rightarrow O}$	3.5×10^{-6}	2.38×10^{-6}	0
	θ_{AC}	55300	θ_O	60300
	$m_{O\rightarrow AC}$	1.55×10^{-9}	$m_{O\rightarrow AC}$	0
	θ_{OCA}	68100	θ_O	117600
Ori+Cas vs. Amaz	Log(p)	-9.0082 -9.2700	-10.6971 -10.9108	-10.2360 -10.2307
	T	1.87	1.78	0.066
	σ of *t*	56900	21700	160
	θ_{OC}	74400	71900	91400
	$m_{A\rightarrow OC}$	3.33×10^{-6}	2.43×10^{-6}	0
	θ_A	62800	θ_{OC}	80200
	$m_{OC\rightarrow A}$	2.17×10^{-9}	$m_{OC\rightarrow A}$	0
	θ_{OCA}	39600	θ_{OC}	120900

Figure 5

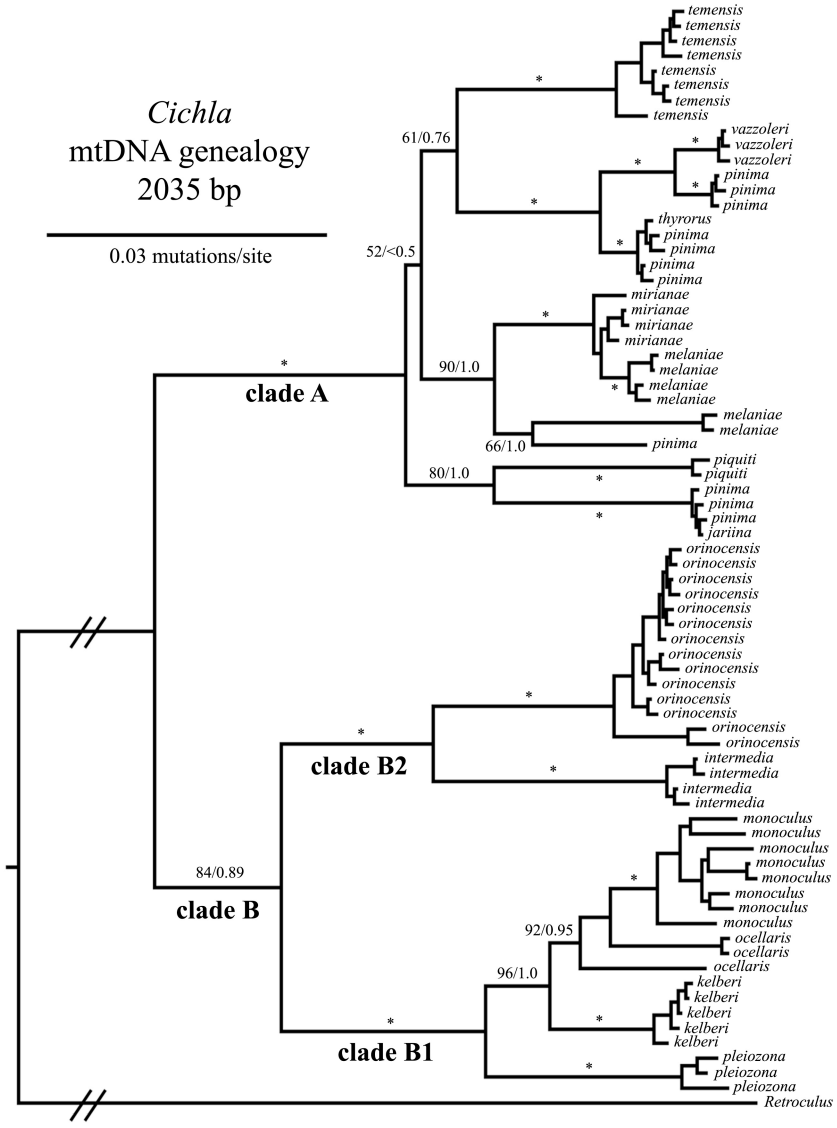
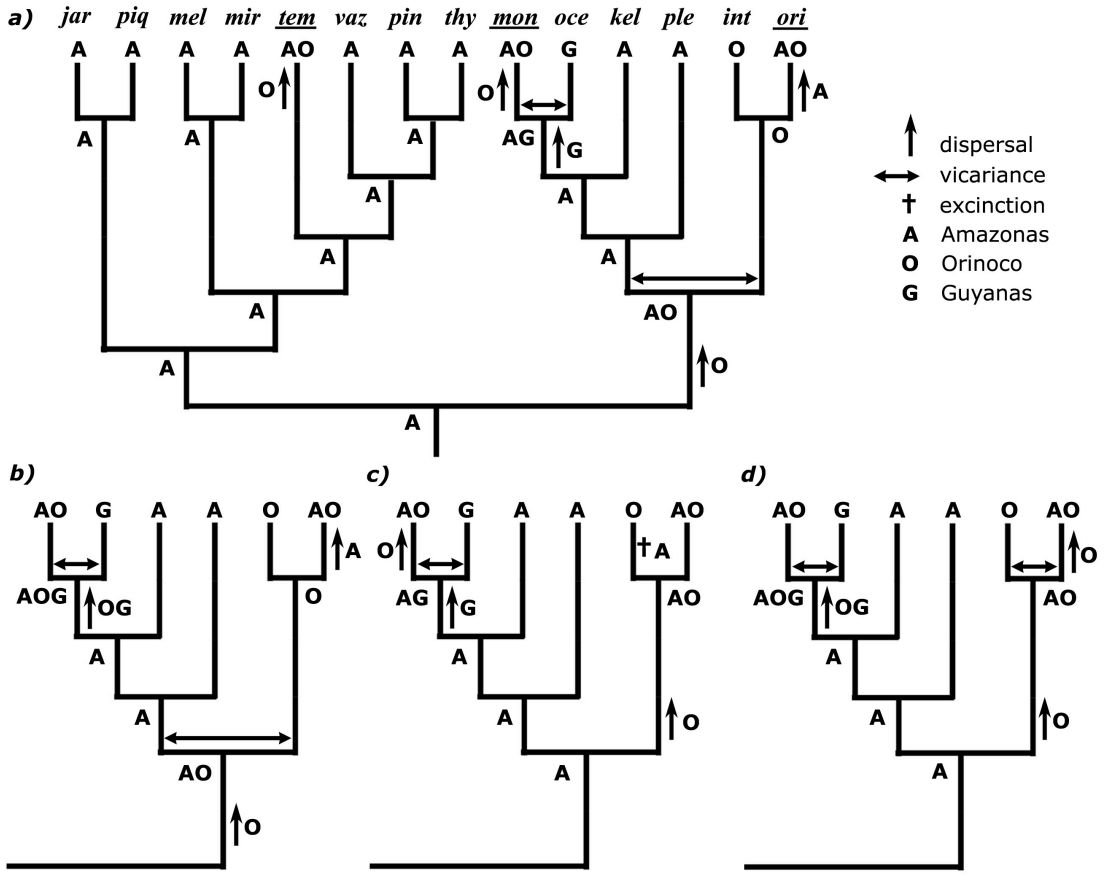

Cichla
mtDNA genealogy
2035 bp
0.03 mutations/site
temensis
temensis
temensis
temensis
temensis
temensis
temensis
temensis
vazzoleri
vazzoleri
vazzoleri
pinima
pinima
pinima
thyrorus
pinima
pinima
pinima
pinima
mirianae
mirianae
mirianae
mirianae
melaniae
melaniae
melaniae
melaniae
melaniae
melaniae
pinima
piquiti
piquiti
pinima
pinima
pinima
jariina
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
orinocensis
intermedia
intermedia
intermedia
intermedia
monoculus
monoculus
monoculus
monoculus
monoculus
monoculus
monoculus
monoculus
ocellaris
ocellaris
ocellaris
kelberi
kelberi
kelberi
kelberi
kelberi
pleiozona
pleiozona
pleiozona
Retroculus
61/0.76
52/<0.5
90/1.0
66/1.0
80/1.0
clade A
84/0.89
clade B
clade B2
clade B1
92/0.95
96/1.0
*

Figure 6


a)
jar piq mel mir tem vaz pin thy mon oce kel ple int ori
b)
c)
d)
dispersal
vicariance
excinction
A Amazonas
O Orinoco
G Guyanas

Chapter 5: Testing mitochondrial capture and deep coalescence in Amazonian cichlid fishes (Cichlidae: *Cichla*)

Authors: Stuart C. Willis, Izeni P. Farias, & Guillermo Ortí

Introduction

Hybridization and introgression are of key interest in evolutionary study. Hybridization can be a critical factor in the process of speciation under the biological species concept because of the strengthening of reproductive isolation via reinforcement (Coyne and Orr 2004). Introgression, where heterospecific DNA is exchanged between species through the backcrossing of hybrids with pure parental types of one or both species, can also have significant evolutionary consequences, ranging from increased adaptive potential and adaptive radiation to extinction by fusion of poorly isolated species (Dowling and Secor 1997; Mallet 2005; Arnold and Martin 2009). Moreover, the exhibition of putatively heterospecific DNA is becoming a relatively common inference for phylogeographic studies (Funk and Omland 2003; Mallet 2005; Mallet et al. 2007). In particular, it has been suggested that the mitochondrial genome (mtDNA) may be particularly prone to cross species boundaries (Chan and Levin 2005), despite the growing evidence for the non-neutrality of mtDNA mutations and co-evolution between conspecific mtDNA and nuclear genomes (Dowling et al. 2008; Galtier et al. 2009). The reason for this may simply be increased genetic drift due to the reduced effective population size of the molecule, owing from its haploid and maternal inheritance in most animals (Ballard and Whitlock 2004), although it has also been observed that positive selection may play a role in accelerating post-hybridization fixation (e.g. Bachtrog et al. 2006).

One important caveat to the inference of introgression is that most studies fail to include adequate tests of alternative hypotheses, principally the stochastic sorting of ancestral allele lineages (Knowles and Maddison 2002). While the identification of instances of introgression are generally based on discordance between gene trees, or between morphology and genetic lineage, stochasticity in the genetic drift of ancestral

polymorphisms is a process well-known for producing similar phenomena (Figure 1; Pamilo and Nei 1988; Maddison 1997). Ancient introgression, when hybridization is no longer contemporary, may be particularly difficult to distinguish because extant species no longer share the same derived alleles, the introgressed alleles may no longer correlate with areas of sympatry, and/or the allele lineages may have become monophyletic following introgression (e.g. Linnen and Farrell 2007). Fortunately, the null hypothesis of lineage sorting makes predictions about the age of discordance events, specifically that deeply coalescing lineages must precede the divergence (speciation) event of the species in which they are found (Pamilo and Nei 1988; Takahata 1989; see also Edwards and Beerli 2000; Degnan and Rosenberg 2009). Using time-calibrated genealogies, it is possible to test this assumption. Further, for neutral polymorphisms, the lineage sorting process is governed by the time interval between speciation events and the effective size of the population at that time. Shorter time intervals and larger population sizes tend to produce a higher degree of discordance among gene lineages, and between gene trees and the organismal phylogeny (i.e. species tree) (Pamilo and Nei 1988; Nichols 2001). Using the pattern of coalescence among multiple genes, it is possible to estimate these parameters for extant and ancestral species (Rannala and Yang 2003) and estimate the probability of observed topological patterns (Degnan and Salter 2005). In true instances of introgression, the inclusion of introgressed alleles can be expected to alter the estimated parameters of a multispecies coalescent model, potentially by increasing the inferred ancestral effective population size or decreasing the time interval between speciation events in order to accommodate more discordance among genes.

We used these predictions to test the discordant mtDNA pattern observed in a widespread species of Neotropical cichlid, the butterfly peacock bass (*Cichla orinocensis* Humbolt, 1821). This species is found in the Orinoco River basin and the Negro River, a major tributary of the Amazonas River basin; these two drainages, which flow separately to the Atlantic, are connected by a unique biogeographic corridor: the Casiquiare connection (Winemiller and Willis 2010). We previously observed that the mtDNA in this species was largely divided into two geographically adjacent clades: one distributed in the Orinoco and upper Negro and the other in the lower Negro River (Figure 2). More importantly, these clades were polyphyletic, that is, phylogenetically resolved with two

separate groups of species: a smaller clade with *C. ocellaris s.l.* (clade B1) and the more populous and widespread clade with *C. intermedia* (clade B2), and we hypothesized that this pattern resulted from an ancient introgression (mitochondrial capture) event (Willis et al. 2007). In our recent study of species boundaries in this genus, we further observed that although these mtDNA clades were only found together at only a single sampling locality, microsatellites portrayed this species as a continuous, interbreeding meta-population, albeit one likely characterized by isolation-by-distance among its subpopulations (Chapter 2). This was also supported by morphological data (Kullander and Ferreira 2006). We also observed that *C. orinocensis* exhibited ongoing introgressive hybridization with both of the species to which its mtDNA was related (clade B1+B2), although this was numerically rare in terms of individuals (Chapter 2). In contrast, we observed that in a concatenated multilocus phylogenetic analysis, *C. orinocensis* was resolved as monophyletic and sister to *C. intermedia* (clade B2) (Chapter 3). While these observations make ancient introgression a likely explanation for the mtDNA topology in *C. orinocensis*, the large (census) population sizes and unknown degree of discordance in the nuclear gene trees makes incomplete lineage sorting another possible explanation. However, using coalescent-based multilocus species tree analyses and genealogical simulations, in addition to dated comparisons of genealogical expectations, we show that the retention of ancestral polymorphism is not a plausible source for the mtDNA genealogy in this species.

Methods

Matrix construction

We obtained sequences of one mtDNA (ATPase 8,6) and twenty nuclear loci collected in our previous studies (Willis et al. 2010; Chapter 2,3) from Genbank for 4 *Cichla orinocensis* (2 from each mtDNA clade), 2 *C. intermedia*, 3 *C. ocellaris* sub. *monoculus*, and 1 *C. temensis* (outgroup) (Accession #s XXXXXXX-XXXXXXX) (Table 1). These individuals were chosen to avoid the influence of recent introgression (for sampling localities, see Chapters 2,3). The loci included introns, contiguous exon and intron regions, microsatellite flanking regions, and anonymous loci and exhibited no signs of paralogous copies among sequences (Chapter 3). As several of our analyses are

coalescent-based and expect mutational distributions derived from haplotypes, we phased our nuclear loci using the Bayesian MCMC program PHASE (Stephens et al.) under the recombination model (MR). In addition, as coalescent methods are very sensitive to accurate estimates of branch lengths and generally assume that sites within a locus are strictly linked (i.e. no recombinants), we tested for recombination among ingroup haplotypes of nuclear loci using the four gamete test (Hudson and Kaplan 1985) implemented in DNASP v5.1 (Rozas et al. 2003). Most loci exhibited no evidence of recombination, but for several loci we trimmed the sequences to retain the most variable segment of DNA that did not exhibit evidence of recombination. In order to pair the mtDNA haplotype with the nuclear sequences for species tree analyses, we randomly chose one of the two alleles for each individual from each of the twenty loci.

Tests of genealogical affinity

Although the maximum likelihood estimate of the mtDNA genealogy resolved the *C. orinocensis* mtDNA as polyphyletic (Chapter 2), one conjecture is that this is an artifactual result stemming from noise in the phylogenetic signal rather than the true pattern. In order to test this, we made heuristic searches in PAUP* v 4b10 (Swofford 2002) where the clade B1 *C. orinocensis* were constrained to group with *C. oc. monoculus*, constrained to group with *C. intermedia*, or not constrained. Searches consisted of 10,000 random-addition sequence replicates limited to 10,000 topological rearrangements per replicate. We compared the resulting trees with a Shimodaira-Hasegawa topology test of significance (Shimodaira and Hasegawa 1999) in PAUP* using 1000 RELL bootstrap replicates. The model of evolution was one chosen for a larger dataset of ATPase sequences (N=120, Chapter 2) in MRMODELTEST using the Akaike Information Criterion (AICc; Akaike 1974), including parameter estimates made there (Table 1).

Another important question regarding this topology is how does it compare to the distribution of topologies among genes; that is, do the nuclear loci exhibit a similar polyphyletic pattern resulting from introgression or incomplete lineage sorting? To test this, we made heuristic searches in PAUP* for each of the 20 loci under the two topological constraints. Searches were performed as above using models estimated in

MRMODELTEST for the full phylogenetic dataset for each locus (Chapter 3; Table1). Resulting trees were compared using a Kishino-Hasegawa topology test (Kishino and Hasegawa 1989) using 1000 RELL bootstrap replicates in PAUP*.

Species tree estimation

We used a Bayesian estimate of species tree, BEST (Edwards et al. 2007; Liu et al. 2008), to simultaneously estimate the species tree, including topology, branch lengths, and effective population sizes for extant and ancestral species and its constituent gene trees. This program, a modified version of MRBAYES, implements a multispecies coalescent model (Rannala and Yang 2003) to search for the optimal species tree and gene trees using MCMC. While this analysis does not make a strict assumption of a molecular clock on mutation rates for each locus, it does scale proposed branch lengths to satisfy the molecular clock assumption of the coalescent model (Liu and Pearl 2007). In any event, we previously tested the molecular clock assumption for these loci and found that none failed the molecular clock hypothesis (Chapter 3). We made two separate runs using only the nuclear data and excluding the sequences from the clade B1 *C. orinocensis* to avoid any influence of introgression in the species tree. The number of alleles for each locus, including all species, was thus 8. Each run consisted of 20 chains with a heating value of 0.05, a length of 200 million generations with a sampling frequency of 5000, and a burn-in of 100 million generations, resulting in 20,000 samples per run. Models for each locus (partition) were estimated for this dataset using MRMODELTEST (Table 1), and the *trratio* prior for applicable partitions was set to approximate the empirical ratio of transitions and transversions by using a beta distribution with alpha and beta equal to the observed number of each mutation type plus one as recommended by the MrBayes website (http://mrbayes.csit.fsu.edu/Help/prset.html). Because of the overall low variability of the nuclear data, the *brlenspr*, the prior on branch lengths, was given an exponential distribution with a rate parameter of 1000 (i.e. mean of 0.001). The *GeneMuPr*, the prior that describes the variation in relative mutation rates among loci, was given a uniform distribution with limits of 0.1 and 2, thus allowing for a twenty-fold variation in rates. The *poissonmean* variable that describes the neighborhood size (# nodes) around the MT tree used to propose changes in the MCMC chain was set to 2. The

propTemp parameter that determines the number of generations over which the chain is cooled from its initial heated state was set to 0.2 (20%). Finally, the prior on effective population size, *thetapr*, was given an inverse gamma distribution with alpha of 4 and beta of 0.003. These parameters were chosen to approximate the average Θ estimated using Watterson's estimate of per site nucleotide polymorphism (Watterson 1975) calculated in DNASP from alleles of the ATPase, *mitf* and *xsrc* loci from several individuals per ingroup species (Chapter 2). MtDNA Θs were divided by two before averaging to account for haploid inheritance (http://www.stat.osu.edu/~dkp/BEST/help/). Convergence of each run to a stable posterior distribution was estimated using Tracer (Rambaut and Drummond 2007) by observing the plots of posterior probability over time.

Concordance of clade divergence times

The null hypothesis of incomplete lineage sorting in *C. orinocensis* mtDNA suggests that the divergences in extant mtDNA lineages were contemporary or preceded the divergences in extant species. In order to be consistent with this hypothesis, the divergence of *C. orinocensis* clade B1 mtDNA from *C. oc. monoculus* (T_{B1} of Figure 1) must be as old as or older than the divergence of clade B1 and B2 species (τ_{cladeB} of Figure 1). To test this conjecture, we estimated a time-calibrated genealogy of the mtDNA using BEAST (Drummond and Rambaut 2007). To calibrate the tree, we used a crown (root) age for *Cichla* of 16.6 million years ago (MYA) estimated from a larger time-calibrated phylogeny of Neotropical cichlids based on three fossils and the break up of Gondwana (López-Fernández et al. unpublished). While secondary calibrations are generally to be avoided (Graur and Martin 2004; Ho and Phillips 2009), we incorporated uncertainty in this calibration (95% highest posterior density, HPD, or the shortest interval that contained 95% of the posterior sample: 7.9-26.3 MYA) into our test. In order to be as consistent as possible with the Yule speciation prior, which assumes a single tip per species, we used a single sequence from each of the ingroup and outgroup species and one from each of the *C. orinocensis* mtDNA clades (five taxa total).

We made two separate analyses using these conditions. The first was designed to estimate the mean age of the mtDNA clades. This analysis used a normal prior for the

treemodel.rootheight parameter with a mean and initial value of 16.6 MYA and a standard deviation of 0.1 and implemented a strict molecular clock on mutation rates (see below). The topologies of sampled trees were constrained to have the ML topology. Other priors were left as default. The second analysis was designed to estimate the uncertainty in the age of mtDNA clades in order to test congruence with the species tree. This second analysis used a relaxed (uncorrelated lognormal) molecular clock on mutation rates (Drummond et al. 2006) with normal priors on the *ucld.mean* prior (average mutation rate across branches) and *meanRate* prior (average of branch lengths divided by time intervals) with means of 0.003 for both, and standard deviations of 0.003 and 0.002 respectively. The *treemodel.rootheight* prior was given a normal distribution with a mean of 16.6 MYA and a standard deviation of 6. This prior gives a 95% HPD on root ages of 6.7-26.4 MYA, approximating the uncertainty in the original time-calibrated analysis (López-Fernández et al. unpublished). Each analysis was run twice for 100 million generations with a sampling frequency of 1000 generations, and the first 10 million were treated as burn-in. We implemented a TN+Γ model (estimated with MRMODELTEST for these sequences) with the Yule speciation prior and a UPGMA starting tree. The resulting 95% HPD intervals for the divergence times of the mtDNA clades were compared to the mean estimates of the organismal (nuclear) phylogeny to test for congruence. To examine the influence of prior values on posterior estimates of divergence times, we ran an analysis without data (empty matrix) for 10 million generations and compared posterior divergence time estimates with those from the analysis with data.

The assumption of a molecular clock for the nuclear loci was examined earlier (Chapter 3), but for the mtDNA locus, this was done here in two ways. First, we observed the *coefficientOfVariation* parameter in the relaxed clock analyses that described the variation in rates among branches. It has been suggested that if the posterior distribution of this parameter abuts zero, then the data cannot reject a strict molecular clock (Drummond et al. 2007). To test the strict clock assumption, we observed the posterior distribution of this parameter from the relaxed clock analyses. Second, we compared the likelihood scores of the mtDNA model with and without the molecular clock assumption in PAUP* using a likelihood ratio test (Huelsenbeck and Bull 1996).

Divergence times for the organismal phylogeny were estimated in two ways. First, we made a BEAST analysis of the 20 nuclear loci (one allele per species per locus) using a separate mutation model and strict molecular clock for each locus, a unified tree prior, and a normal prior for the *treemodel.rootheight* parameter with a mean of 16.6 MYA and a standard deviation of 0.1. Two runs were made as above; however, we used a user-specified starting tree rather than UPGMA. This tree exhibited the topology from our previous concatenated analysis (Chapter 3). The root age was specified as 16.6 MYA, the divergence of *C. orinocensis* and *C. intermedia* was specified as 0.5 MYA, and the divergence of *C. oc. monoculus* and the ancestor of the previous species was specified as 1.0 MYA. Because the mtDNA divergences are expected to be older than the organismal (nuclear) divergences under a hypothesis of deep coalescence, we considered this a conservative starting tree. We also ran a single analysis without data to examine the influence of the priors, as above. Second, we transformed the node heights from the BEST species tree analysis using the mean root age of 16.6 MYA via TreeEdit v1.0 (A. Rambaut, University of Oxford).

Species tree using introgressed DNA

The coalescent models used in BEST estimate the branch length and population size parameters of the model based on the distribution of coalescence among gene trees (Rannala and Yang 2003; Liu and Pearl 2007). This analysis assumes that the only process creating topological discordance or variation in coalescence times among loci is the lineage sorting process itself, that is, genetic drift within lineages (Liu and Pearl 2007). Incorporating gene trees that have topologies resulting from introgression, horizontal gene transfer (e.g. viral), or paralogy could alter the parameter estimates for the model, including the species tree topology, branch lengths, and population sizes (for example, see Eckert and Carstens 2008). For comparison to the species tree using only nuclear data and the clade B2 *C. orinocensis*, we ran three other analyses with identical run conditions, with the exception that the mtDNA locus was specified as haploid and the *GeneMuPr* was set to {0.1,10} to accommodate the mtDNA mutation rate. First, we ran an analysis that included both the clade B1 *C. orinocensis* and the mtDNA locus. Next, for comparison to the results of these two species tree analyses, we ran an analysis that

only used the nuclear data but included the clade B1 *C. orinocensis* and another analysis that included the mtDNA but did not include the clade B1 *C. orinocensis*. These latter analyses were run to separate the influence of the clade B1 *C. orinocensis* nuclear DNA from their mtDNA and the inclusion of the more variable mtDNA data itself from the clade B1 *C. orinocensis* alleles. We expected that in the analysis with both *C. orinocensis* mtDNA groups, if our alternative hypothesis of introgression was correct, then the parameter estimates for this species tree would differ from the clade B2-nuclear only tree. In order to accommodate more discordance among loci introduced with introgressed gene tree topologies, the analysis could presumably increase population sizes or decrease speciation intervals. In particular, we expected larger Θ values (population sizes) or smaller branch lengths for the ancestor of *C. orinocensis* and *C. intermedia*, reflecting less genetic drift and lineage sorting in this ancestral population, which would explain discordance introduced with the clade B1 *C. orinocensis* mtDNA alleles. For each of the species tree analyses, we calculated the coalescent branch lengths (speciation intervals in units of generations calibrated to the effective population size; see below) for each tree in the posterior distribution. We examined whether the 95% HPD for each of the ingroup branches included the mean from the clade B2-nuclear only species tree.

Simulation of gene trees under the coalescent

One limitation to the above species tree comparison is that in the BEST analysis the gene trees parameters, topology and branch lengths, are sampled along with the rest of the model. If the information content of some data partitions is limited, the genealogical signal from these loci may be overwhelmed by the signal from other partitions, including those influenced by introgression. The result could be that the model results do not differ, not because the data are not incongruent, but because incongruence is masked by weak signal from several partitions. An alternative test of the probability of the topology of the *C. orinocensis* mtDNA gene tree would be one where the topology and branch lengths were known rather than estimated (Buckley et al. 2006). Thus, we simulated gene trees based on the parameters estimated in the species tree without the influence of introgression and examined the frequency of the observed mtDNA topology

among these genealogies. We considered the hypothesis of lineage sorting unlikely if the probability of the mtDNA topology was less than 5% (Buckley et al. 2006).

Using MESQUITE v2.6 (Maddison and Maddison 2008), we simulated 10,000 gene trees constrained to evolve only through lineage sorting in an organismal phylogeny ("Coalescence contained within current tree"). In each gene tree, we simulated a single allele for *C. temensis*, *C. oc. monoculus*, and *C. intermedia*, and 10 alleles for *C. orinocensis*. The species tree used in MESQUITE reflected the tree topology, branch lengths, and population sizes estimated using BEST from only the nuclear data and clade B2 *C. orinocensis*. We input into MESQUITE a species tree with branch lengths (time) in units of generations and branch width (effective population size, N_e) multipliers in units of 100K gene copies. As MESQUITE simulates *haploid* individuals (gene copies), we calculated the number of gene copies in the diploid population using $\Theta = 2N_e\mu$ (see below). We made four analyses using the same topology and branch lengths, but different population sizes. The first, simulating coalescence in the nuclear genes, used population sizes converted from the mean Θ values estimated for extant and ancestral species using BEST. However, it has been suggested that ancestral population sizes are among the most difficult parameters to estimate accurately using these types of programs (Edwards et al. 2007; Heled and Drummond 2010) and may be prone to underestimation when there is insufficient information among loci to estimate coalescence in the ancestral populations (Wakeley and Hey 1997). Therefore, we also simulated nuclear gene trees using ancestral population sizes that were predicted as the average of the branches they subtended (i.e. their daughter populations). Next, we simulated topologies for the mtDNA gene tree using both of the aforementioned sets of population sizes, but divided each by four to reflect the reduced effective population size (effective number of gene copies) of the haploid and maternally inherited mtDNA genome (Palumbi et al. 2001; Ballard and Whitlock 2004). For each of the four sets of gene trees, we used topological filters in PAUP* to estimate the proportion of trees in which *C. oc. monoculus* was not sister to *C. orinocensis*+*C. intermedia* (that is, for unrooted topologies, the proportion of trees in which *C. temensis* and *C. oc. monoculus* were not sister), and of these, the proportion in which *C. orinocensis* was not monophyletic, reflecting the observed mtDNA topology

(although this includes trees in which *C. oc. monoculus* may be nested among *C. orinocensis*).

To convert the parameter estimates from BEST into meaningful numbers (generations and individuals), we used the root age of 16.6 MYA and a generation time of 5 years (K. Winemiller, pers. comm.). These calibrations are necessary to calculate the mutation rate in units of mutations/site/generation (see below). However, in actuality, these estimates were only a convenience for use in MESQUITE. Coalescence is expected to be the same for branches where the ratio of branch length (generations) and population size (individuals) are the same, that is, equal when measured in units of N_e generations, because the rate of coalescence in a population depends on this ratio (Kingman 1982). Thus, we considered using different generation times (e.g. 2 yrs) or younger divergence times (e.g. 7.9 MYA) to estimate the probability of the mtDNA topology given uncertainty in our demographic values, but in each case, the coalescent units are the same because of the effect on mutation rate (mutations/site/generation). This can easily be demonstrated algebraically (see also Edwards and Beerli 2000; Degnan and Rosenberg 2009). For example, ultrametric branch lengths in units of mutations/site can be converted to generations using the following formula,

$$\text{Generations} = \text{mutations/site} \div \text{mutations/site/yr} \div \text{yr/gen}$$

or

$$\text{Generations} = \text{mutations/site} * \text{gen/mutation/site},$$

where yr/gen is the generation time and mutations/site/yr is the standard per site mutation rate. Similarly, the Θ values associated with branches can be converted to estimates of diploid individuals using the standard formula $\Theta = 4N_e\mu$, or to estimates of the effective number of gene copies using $\Theta = 2N_e\mu$, where μ is the mutation rate per site per generation (Kuhner et al. 1995). This can be rearranged as

$$N_e = \Theta \div 4\mu,$$
$$N_e = \Theta \div (4 * \text{mutations/site/yr} \div \text{yr/gen}),$$

or

$$N_e = \Theta \div (4 * \text{gen/mutation/site}).$$

Finally, a branch length in coalescent units can be calculated by dividing the branch length in generations by the population size in gene copies to give generations in multiples of N_e, which appears for diploid individuals as

$$\text{Generations} / N_e = \frac{\text{mutations/site} * \text{gen/mutation/site}}{\Theta \div (2 * \text{gen/mutation/site})},$$

and, when mutation rate is factored out, this results in

$$\text{Generations} / N_e = 2 * \text{mutations/site} \div \Theta,$$

or two times branch length divided by Θ (Degnan and Rosenberg 2009). Thus, because mutation rate is factored out in the end, and any change in divergence time or generation length is ultimately carried through this parameter, changes of this sort have no effect on our estimate of genealogical patterns as a result of coalescence.

Results

The nuclear loci exhibited relatively few variable positions after accommodating potential recombination, particularly for the ingroup taxa and in comparison to the mtDNA locus (Table 1). Perhaps as a result, the models of evolution recommended for these nuclear loci were all relatively simple, with the most complex being the six-parameter HKY+I model. Potentially also due to low variability, the only gene topology that was significantly incongruent with the constraint for all *C. orinocensis* to group with *C. intermedia* (i.e. the concatenated nuclear topology; Chapter 3) was the mtDNA tree ($p = 0.006$; Table 1). As expected, the unconstrained maximum likelihood topology was recovered under the polyphyletic constraint, where the clade B1 *C. orinocensis* had to group with *C. oc. monoculus*, so an SH test was used to compare the ML and constrained topologies. In contrast, although two topologies for the nuclear loci had higher likelihood scores under the polyphyletic constraint, neither of these was significant under a KH test. Moreover, the only locus with a significant difference for these topologies favored the concatenated topology (Table 1).

We observe that the age of the clade B1 node in the mtDNA tree is not consistent with a divergence prior to the divergence of the clade B species in the organismal tree (Figure 3). The posterior distribution for the species tree in the BEST analysis contained only a single tree (i.e. 100% posterior support for all nodes after burn-in). This tree was congruent with the topology of the concatenated nuclear tree we previously inferred (Chapter 3). A time-calibrated version of this tree based on a crown age estimate of *Cichla* of 16.6 MYA (López-Fernández et al. unpublished) is shown in Figure 3. Notably, the age of the first speciation in clade B (τ_{cladeB} from Figure 1) was estimated at 12.2 MYA (95% HPD of 9.9-14.4) (Table 2). In contrast, when the basal divergence in the mtDNA tree was calibrated to 16.6 MYA using BEAST, the mean age for the divergence of clade B1 *C. orinocensis* and *C. oc. monoculus* (T_{B1} from Figure 1) was 1.5 MYA in the strict clock analysis. The 95% highest posterior density interval (HPD) of the age of this node, made using a relaxed molecular clock and incorporating uncertainty in the root calibration, was 0.2 to 3.5 MYA (Figure 3, Table 2). The likelihood ratio test using likelihood scores estimated with PAUP* failed to reject the molecular clock hypothesis for the mtDNA tree (2*diff = 7.43738, p = 0.0591 with 3 d.f.), and the coefficient of variation for the relaxed clock analysis in BEAST abutted zero, implying that the data could effectively be modeled with the strict clock assumption. When we estimated a time-calibration for the organismal phylogeny using a partitioned analysis of the 20 nuclear loci in BEAST, the posterior estimates of nodal ages were somewhat different from the BEST species tree (Table 2). These age estimates were much more similar to those estimated from the prior distribution without data, suggesting the Yule speciation prior had an undue influence on this calibration. In contrast, the age estimates for the mtDNA tree were quite different from their prior distributions.

Each of the species trees with clade B1 *C. orinocensis* and/or mtDNA in BEST favored a single topology in the posterior distribution, and this topology was the same as observed above with neither of these data (i.e. the concatenated topology, Figure 3). However, the parameters estimated for some analyses differed. Incorporation of the clade B1 *C. orinocensis* into the BEST analyses caused a decrease in the coalescent branch lengths for *C. orinocensis* in those two species trees, but only in the analysis with all data (clade B1 *C. orinocensis* AND mtDNA) was there a significant decrease in the coalescent

branch lengths for the ancestor of *C. orinocensis* and *C. intermedia*, and the ancestor of these two and *C. oc. monoculus* (Table 3). The decrease in this analysis reflected an increase in discordance among loci when the clade B1 mtDNA alleles were added relative to the other data combinations. The difference in coalescent branch lengths estimated by BEST appears to derive from a reduced speciation interval in the former ancestor and an increased population size (Θ) in the latter ancestor (Supplemental Tables 1 and 2). Another pronounced difference was the increased root-to-tip depth of the species trees including mtDNA (0.00656 and 0.00612 versus 0.00436 and 0.00411 for those without, Suppl. Table 1), reflecting the higher mutation rate of the mtDNA locus relative to the nuclear loci (Table 1). Similarly, the range of relative mutation rates was from 0.666 to 1.304 (ratio 1.95) and 0.52 to 4.839 (ratio 9.31) in the nuclear and nuclear plus mtDNA analyses respectively. This also means that the range of mutation rates should have been sufficiently encompassed by the *GeneMuPr* prior in each case (see Methods). Interestingly, although the posterior tree lengths for the mtDNA partition in the two analyses with this locus were very similar (mean TL=0.110, 95% HPD 0.087-0.133 without clade B1 *C. orinocensis* and mean TL=0.118, HPD 0.093-0.146 with: a difference of 7%), the relative mutation rate for this partition was more than 20% lower in the analyses with clade B1 *C. orinocensis* (mean rate 4.84, HPD 4.14- 5.73 versus mean rate 3.80, HPD 2.86-4.45, respectively). The same tree length with a slower rate implies a longer tree relative to the root-to-tip distance of the species tree and other genealogies, and this too may reflect the analysis trying to accommodate greater discordance under a strict model of deep coalescence.

Based on the first species tree from BEST calibrated to a root age of 16.6, the estimated mutation rate for the nuclear loci is 2.63×10^{-10} mut/site/yr (95% HPD of $1.38 \times 10^{-10} – 6.38 \times 10^{-10}$) or 1.31×10^{-9} mut/site/gen for a 5 yr generation time. Not surprisingly, this was significantly slower than the ATPase rate, estimated using BEAST, of 3.36×10^{-9} mut/site/yr (HPD of $2.3 \times 10^{-9} – 4.5 \times 10^{-9}$). We used this nuclear mutation rate (per generation) to estimate population sizes for the coalescent simulations with MESQUITE, along with a root age of 3.32×10^{6} generations (16.6 myr ÷ 5 yr/gen). The simulation of genealogies under different population sizes adjusted for nuclear and mtDNA N_e provided strikingly different results (Figure 4). For example, using the

averages of population sizes for extant species supplied by the BEST analysis as the ancestral population sizes, approximately 37% of the simulated genealogies included topologies in which the 10 *C. orinocensis* alleles were both not monophyletic and not sister to *C. intermedia*. When genealogies for the mtDNA were simulated under ¼ of this population size, we observed only 103 topologies out of 10,000 that exhibited these characteristics (1%). Moreover, when we used the ancestral population sizes estimated directly with BEST, we found that with the unreduced (nuclear) population sizes, 12% of the simulated genealogies portrayed *C. orinocensis* as para- or polyphyletic and sister to a species other than *C. intermedia*, but none were recovered under the empirical mtDNA parameters. In general, *C. orinocensis* was much more likely to be non-monophyletic with respect to other species for the genealogies simulated under the full population sizes. For the mtDNA, we inferred that the probability of observing a discordant topology under the hypothesis of deep coalescence was 0.01 or less by taking the simulation results directly, as the mitochondria is unique in the genome (i.e. there is only one linked chromosome to sample, wherein all loci have the same genealogy). To estimate the probability of observing two or more discordant nuclear gene trees, as we potentially did according to the topology test results (Table 1), we used the estimated individual probabilities (0.1194 and 0.3678) in a binomial calculation with a sample size of 20. This predicts that the probability of observing this two or more discordant gene trees under these conditions would be 0.708 and 0.999, respectively.

Discussion

Distinguishing deep coalescence and introgression

The discordance of gene trees from a sampled set of species has several potential origins, including horizontal gene transfer, introgressive hybridization, unrecognized paralogy, and deep coalescence (Maddison 1997). However, unlike the other phenomena, deep coalescence is a neutral process that happens without the movement of genes within or between genomes. Moreover, it can be predicted by the demographic and phylogenetic conditions of a lineage, most of which are estimable from multilocus genetic data, and this process therefore serves as a null model for gene tree discordance (Degnan and Rosenberg 2009). For example, unlike deep coalescence, a hypothesis of recent or

ongoing introgressive hybridization would suggest that gene tree discordance should correlate with areas of sympatry between putatively hybridizing species (Donnelly et al. 2004; Morando et al. 2004). However, where introgression has been occurring for a long period of time, or when hybridization occurred some time in the past, this prediction may no longer apply (Linnen and Farrell 2007; Peters et al. 2007). On the other hand, deep coalescence always requires that allele lineages diverge prior to the species in which they are found; introgressed lineages may coalesce much more recently than the species themselves (Degnan and Rosenberg 2009). When sequences from multiple unlinked loci are available, it is possible to test for bimodality in gene divergences resulting from vertical (speciation) and horizontal (introgression) processes (e.g. Peters et al. 2007; Good et al. 2008).

Based on our results, the mtDNA genealogy of *Cichla orinocensis* is inconsistent with a model of incomplete lineage sorting or deep coalescence, and most likely derives from a history of introgression and mitochondrial capture. Based on the dating of the mtDNA genealogy, this introgression event occurred around 1.5 MYA, well after the divergence of the clade B species as estimated from nuclear data (12.2 MYA) (Table 2). However, two things should be taken into consideration with this result. First, we experienced some difficulty in stably estimating organismal divergence times from the nuclear data. For example the dates estimated by BEAST using the Yule speciation prior were quite different from those derived from the BEST tree. Rather, the former were much more like those estimated from the model without data, suggesting that the prior had an undue influence on this analysis despite using strict molecular clocks on each locus. Estimates made with a birth-death speciation prior exhibited similar dependence on the prior distribution (not shown). In a salient study, Brown and Yang (Brown and Yang 2010) showed that the dating of young phylogenies is likely to be particularly difficult because of the low degree of divergence among species. These authors suggested that without the incorporation of rapidly evolving loci (mutation rates >0.01 mut/site/myr), analyses of young phylogenies with relaxed clocks appear to be highly susceptible to prior influence. While our divergence time estimates were based on a considerable amount of sequence data (10.9 Kbp in total) and strict molecular clocks, our loci were anything but rapidly evolving (~0.00026 mut/site/myr). In contrast, we found that the

ultrametric tree estimated by BEST under a multispecies coalescent model had relative branch lengths much more similar to the total dataset analyzed with a partitioned MRBAYES analysis (Chapter 3). We did not explore the use of the multispecies coalescent model in BEAST (i.e. *BEAST; Heled and Drummond 2010), but it is possible this would perform similarly well. Nevertheless, as this tree was already ultrametric and accounted for the variance between speciation time and gene divergences (e.g. Edwards and Beerli 2000), it was a simple matter to time calibrate this tree. However, second, it was somewhat circular to use the same root calibration for both the organismal (nuclear) phylogeny and the mtDNA genealogy and then compare the divergence of the latter with the former. It is possible, for instance, that the basal divergence in the mtDNA genealogy significantly preceded that of most nuclear genes and that the other divergences in the mtDNA gene tree are also proportionally older. This could potentially allow the divergence of the clade B1 mtDNA (*C. orinocensis* B1 and *C. oc. monoculus*) to be older than the 1.5 MYA estimated for this node. However, that this would change our inference of temporal incongruence seems unlikely for three reasons. For one, if the mtDNA divergences were much deeper in time, meaning that the mtDNA genealogy was constrained only by genetic drift in the MRCA of *Cichla* and possibly of clade B, it would be unlikely for the divergence times between other nodes in the mtDNA and BEST organismal trees to correspond so well (e.g. clade B and *C. orinocensis* B2+*C. intermedia*; Table 2). For another, our relaxed clock BEAST analysis incorporated a great deal of temporal uncertainty regarding the calibration of the root node, and although this uncertainty seems to have been proportional to node age in the posterior distribution, the credible interval was far from including the node age of clade B from the species tree (Table 2). Finally, while there will often be a great deal of variance between gene divergence times and speciation ages (Edwards and Beerli 2000), genes with a smaller effective population size and greater coalescence rate, such as the haploid and maternally inherited mtDNA genome, are expected to more closely follow the organismal phylogeny in the absence of other processes (Palumbi et al. 2001).

In addition to these temporal comparisons, our estimates of species trees, which do not rely on absolute time calibrations, showed a similar incongruence with the expectations of deep coalescence for the *C. orinocensis* mtDNA (Table 3). We found that

when the mtDNA of clade B1 *C. orinocensis* were incorporated into these multispecies coalescent analyses, which assume that all gene tree discordance originates strictly from genetic drift and incomplete lineage sorting, the species tree became significantly shorter (in coalescent units) for the ancestors of clade B species. This was not true for other analyses building upon the species tree using only clade B2 *C. orinocensis* and nuclear data, and implies that these particular mtDNA alleles are discordant with the remaining dataset, as might be expected from introgressed DNA. It is intriguing, in light of the previous discussion of mutation rates and dating, that despite the low variability of the individual loci examined here, as a group these loci were able to represent a cohesive coalescent pattern with which the aforementioned mtDNA alleles were incongruent. A significant amount of discussion over species trees has centered on number of loci necessary to infer a robust species tree (Edwards et al. 2007) or the tradeoff between individuals versus loci (Maddison and Knowles 2006), but relatively little has been said about the information content and value of more, but less variable, loci (but see Lee and Edwards 2008; Huang and Knowles 2009; and Chapter 3). Previous research that successfully used coalescent models to distinguish introgression from deep coalescence used fewer, but significantly more variable, loci (Linnen and Farrell 2007; Peters et al. 2007; Good et al. 2008). The present results suggest that given a sufficient number of loci, even low variability loci, detecting deviations in the overall coalescent pattern, such as those introduced by introgression, is possible. One interesting thing to note was that this pattern appeared in our data despite the mtDNA locus being specified only as haploid (~1/2 nuclear N_e) rather than haploid and maternally inherited (~1/4 nuclear N_e) in the BEST analyses (Liu and Pearl 2007). We expect that if the program accommodated this data characteristic, we would see even greater differences in the results.

In order to estimate the species tree for these taxa and examine the influence of *C. orinocensis* mtDNA lineages, we constructed our datasets without including other representatives of the *Cichla ocellaris sensu lato* species group (Figure 1). While we previously inferred that through evolutionary time the populations within this group have exchanged genes and show apparently low levels of reproductive isolation, each, with the exception of *C. oc. nigromaculata*, has one or more unique mtDNA lineages (Chapter 2). The observation that *C. orinocensis* clade B1 mtDNA is sister to only one of these, *C. oc.*

monoculus, and the nesting of this clade deeply among the *C. ocellars sensu lato* mtDNA lineages, makes a hypothesis of ancient mitochondrial capture between these species more likely. Deep coalescence should require either that 1) mtDNA divergences among subspecies preceded all population divergence events, and that each subspecies retained ancestral polymorphisms and independently became fixed for the mtDNA lineage that represented the monophyly of the complex, or 2) that the *C. orinocensis* clade B1 mtDNA should have been resolved sister to this species complex rather than within it. However, this should not have affected our results from the species tree analyses, and if anything, inclusion of the other mtDNA lineages would have made our results more extreme. However, we chose to exclude sequences from these populations for two reasons. First, while we considered that these lineages were not necessary in order to test our hypotheses of deep coalescence versus ancient introgression in *C. orinocensis*, they would have added considerable computation time to the analyses. Second, and moreover, we were concerned that topological uncertainty regarding relationships among these populations, or alleles within a *C. ocellaris s.l.* lineage, would interfere with estimation of speciation intervals and population sizes for the branches of interest (the ancestors), leading to larger credible intervals and decreasing the power of the test. In any event, it appears that the influence of unsampled taxa on the inference of species trees is a largely unexplored area in need of further study.

In spite of our success with such low-variability loci, there presumably is a point at which the mutation rate at individual loci is simply too low (or divergences too recent) to sufficiently distinguish hypotheses such as deep coalescence and introgression. For example, in the current dataset, although the estimated coalescent branch lengths would suggest a significant degree of native discordance among nuclear loci, we observed no instances in which our nuclear loci were significantly incongruent with the constraint that all *C. orinocensis* group with *C. intermedia* (Table 1). There are two explanations for this: that the loci whose gene trees appeared to prefer a topology of polyphyletic *C. orinocensis* only did so as a result of mutational noise, and under alternative constraint this noise was relatively weak, or that the individual variability of the loci was so low that there was insufficient power for the test. Nevertheless, this observation brings up an important point, which is that the true topology of the gene trees may not be recoverable

given the mutations available to accurately reconstruct it (see also Huang and Knowles 2009). An alternative to these empirical estimates is the simulation of gene trees whose topologies are known with certainty under the conditions prescribed by the candidate models. For instance, Buckley et al. (2006) used a simulation of genealogies under a multispecies coalescent model to determine the probability that a species would be found in two different topological positions in four unlinked gene trees. Using a similar approach here, we estimated the proportion of time that a topology like the *Cichla* mtDNA genealogy would be observed under the null model of deep coalescence. Although use of these genealogies required some simplification of the observed topologies, i.e. treating species nested within a paraphyletic *C. orinocensis* the same as two *C. orinocensis* clades sister to different species and not considering the placement of multiple alleles from *C. intermedia* or *C. oc. monoculus*, these simulations nevertheless confirmed what we inferred from the species tree analyses: that the probability of the mtDNA topology under these conditions was too improbable to ascribe to deep coalescence alone. In contrast, we observed a significant amount of topological discord (polyphyly and disagreements among genealogies) for effective population sizes corresponding to nuclear genes. It should be noted, of course, that it is difficult to assess whether the mtDNA actually experiences an effective population size on the order of ¼ of the nuclear N_e, and if our simulations were accurately arranged. While selective sweeps could further reduce the effective population sizes among the fully linked mtDNA loci (see below), phenomena such as polygamy could proportionally reduce the nuclear N_e with respect to the mtDNA (Ballard and Whitlock 2004). However, most *Cichla* species are known to be seasonally monogamous and exhibit extensive bi-parental care (S. Willis and P. Reiss, pers. obs.; see also Winemiller 2001), making biases which would inflate the relative mtDNA N_e unlikely.

Hybridization and cytonuclear discord in Cichla

The concordance of our several analyses make the inference of introgression clear but do not provide much insight into either the origin of the introgression or the observation that only the mtDNA shows a significant deviation from its expected pattern. Hybridization has long been considered to be a minor process by zoologists, relegated to

fleeting instances of the breakdown in reproductive isolating mechanisms or reinforcement following secondary contact (e.g. Dobzhansky 1937; Mayr 1942). However, genetic surveys are frequently inferring the exhibition of heterospecific DNA (i.e. hybridization), and estimates suggest that 6-10% of animal species hybridize (Mallet 2005). In a recent study using an extensive mtDNA and microsatellite dataset, we inferred that six out of eight delimited species of *Cichla* (9 of 15 described species) hybridized, although numerically this represented a relatively small overall proportion of individuals (<5%) (Chapter 2). In almost every case, mtDNA haplotypes had been transferred between at least one of the two species involved, although nuclear admixture ranged from extensive to undetectable. In the case of *C. orinocensis* and *C. oc. monoculus*, these two were inferred to be experiencing ongoing hybridization in the southernmost Negro River and the adjacent Preta da Eva River (Figure 1). Nine of 20 *C. orinocensis* individuals from these localities exhibited *C. oc. monoculus* mtDNA (i.e. exact control region haplotypes seen in *C. oc. monoculus* from that and other localities), while only 1 of 13 *C. oc. monoculus* exhibited *C. orinocensis* haplotypes. The introgression at nuclear loci was similarly asymmetric, with only *C. orinocensis* exhibiting significant admixture (~18% of alleles) (Chapter 2). Although this ongoing hybridization could not have been responsible for the clade B1 clade of *C. orinocensis* mtDNA (these haplotypes are monophyletic and significantly divergent from extant *C. oc. monoculus* mtDNA), the incomplete reproductive isolation and asymmetric pattern of introgression between these species provides an explanation for how *C. orinocensis* may have captured the mtDNA of *C. oc. monoculus* some time in the past (1.5 MYA, Table 2).

We found no significant evidence of nuclear introgression in *C. orinocensis* that exhibited clade B1 mtDNA outside of recently hybridizing localities. This was evident by the lack of changes in coalescent branch lengths in the species tree analysis using both sets of *C. orinocensis*, but only nuclear DNA. In fact, the observation that maternally-inherited organellar loci (mitochondrial and chloroplast) introgress more often than nuclear loci seems to be relatively common across animals and plants (reviewed in Chan and Levin 2005), despite the growing evidence that there is often tight coordination between mtDNA and nuclear gene products (Dowling et al. 2008). In addition to

accelerated genetic drift due to the reduced effective population size of these loci, several other theories have been suggested. For example, building upon previous work (e.g. Rieseberg et al. 1996; Wirtz 1999), Chan and Levin (2005) suggested that asymmetrical densities of sympatric species with male-male competition or female choice could facilitate the accelerated introgression of maternally-inherited loci. In these situations, where females mediate mating interactions, females of the more rare species have greater mating opportunities with heterospecifics than males of the same rare species, thereby increasing the proportional opportunities for mtDNA introgression into the more common species relative to nuclear DNA. However, it is unclear from the description of these frequency-dependent models whether the more rare species should also exhibit significant effects of introgression, as males of the less common species are only able to mate with conspecific females, but females could be mating with conspecifics, heterospecifics, or hybrids based on frequency. Specifically, this model could explain why *C. orinocensis* in recently hybridizing localities, which appear to be the more common species based on anecdotal evidence, exhibit introgression at both mtDNA and nuclear loci, while *C. oc. monoculus* exhibit very little at the former and virtually none at the latter. However, it is difficult to know, for instance, if this model provides a better explanation for the observed pattern than a scenario in which hybridization between pure parental individuals is overall infrequent, but once it has occurred, *C. orinocensis* show a lower level of discrimination against F_1 and later hybrid generations than do *C. oc. monoculus* (i.e. asymmetric innate reproductive isolating mechanisms).

In any event, although this model provides an explanation for initial biased introgression of mtDNA, it doesn't explain long-term evolution of these introduced mtDNA lineages. Some researchers have suggested that mtDNA may exceed initial sites of hybridization due to positive selection (Ballard and Whitlock 2004). Although there is mounting evidence that there is significant coordination of genetic pathways involving mitochondrial and nuclear gene products and that infection of heterospecific mitochondria in a conspecific nuclear background generally results in reduced fitness (Dowling et al. 2008), one can plausibly imagine a scenario in which the introgression of closely-related heterospecific mitochondria relieves the compensatory burden of the nuclear genome from mutation load on the non-recombining mtDNA (Ballard and

Whitlock 2004). Distinguishing this scenario from simple genetic drift or population expansion will require comparing the distribution of mutations at mtDNA and nuclear loci from abundant population level samples (e.g. Bachtrog et al. 2006). In a scenario of positive selection, the expectation is for mutations to exhibit a star-like phylogeny and an excess of rare mutations for loci under selection (i.e. high haplotype diversity with most haplotypes related by a polytomy) but not for unlinked loci evolving neutrally (Tajima 1989). Examining the mtDNA topology for clade B1 in Figure 2, this seems unlikely, and if anything, the clade B2 topology is more reminiscent of this pattern than clade B1 (see also Chapter 2). Unfortunately, present sample size for the nuclear loci are too small to compare with this mtDNA pattern, but this provides an intriguing line for future investigation.

Conclusions

It is becoming clear that hybridization and introgression play a more important role in evolution than previously thought (Mallet 2005). Deciphering the contribution of hybridization to genetic and taxonomic diversity will require the use of models that distinguish introgression from other patterns and phenomena such as the retention of ancestral polymorphism. We observed that even given a low variability among sequences, examining many unlinked loci can provide a cohesive distribution of coalescence from which introgressed allele lineages stand out. While these models have not yet been fully developed or reviewed, they provide a fundamental shift in the focus from gene trees to the history of populations and the processes that shape them (Edwards 2008).

Table 1. Loci examined in this study, models of evolution (maximum likelihood and Bayesian analysis), and results from topology tests. [a] number of variable sites for ingroup / ingroup+outgroup sequences; [b] p-value for topology test of constrained topologies: * significant at alpha 0.05, [P] *C. orinocensis* polyphyletic in preferred topology; Note: ATPase topology test is SH; all others are KH.

marker	type	base pairs	S[a] IG/OG	ML model	BA model	topology test p-value[b]
mitf	ORF+intron	743	4/9	HKY	HKY	0.319[P]
xsrc	ORF+intron	481	7/7	HKY+I	HKY	0.274
GnRH3i1	intron	374	3/3	F81	F81	0.047*
Gpd2i1	intron	319	3/3	F81	F81	0.072
1835e6	intron	436	6/7	HKY+I	HKY	0.259
8680e2,3	intron	960	7/20	K2P	K2P	0.713[P]
14867e1	intron	409	8/10	HKY+I	HKY	0.304
18049e2	intron	407	4/4	F81	F81	0.249
35564e5	intron	766	9/15	HKY	HKY	0.08
36298e1	intron	458	3/6	HKY	HKY	1
55305e1	intron	592	6/8	HKY	HKY	0.266
55378e1	intron	807	4/4	JC	JC	1
TmoM27	μ-sat flanking	337	1/2	K2P	K2P	1
CorE7	μ-sat flanking	516	1/6	HKY	HKY	1
CorF12	μ-sat flanking	380	6/7	F81	F81	0.053
CinKA7	anonymous	797	3/5	F81+I	F81	0.998
CmoME5	anonymous	494	10/11	HKY	HKY	0.076
CmoMJ3	anonymous	504	4/4	JC	JC	0.101
CteOA7	anonymous	434	7/7	JC	JC	0.071
CteOI2	anonymous	687	9/9	HKY+I	HKY	0.147
ATPase 8,6	ORF (mtDNA)	842	70/98	GTR+I+Γ	TN+Γ	0.006*[P]

Table 2. Divergence times for the mtDNA genealogy and organismal (nuclear) phylogeny estimated using BEAST and BEST. Times are given in millions of years.

Phylogeny	Analysis	Harmonic mean LnL	Root Prior (mean±st.dev.)	Root posterior	MRCA clade B	MRCA clade B2	MRCA clade B1
mtDNA	BEAST (strict clock)	-1647.279	16.6±0.1	16.6 (16.4-16.8)	12.8 (9.6-16.4)	7.0 (4.2-10.0)	1.5 (0.5-2.5)
mtDNA	BEAST (relaxed clock)	-1646.934	16.6±6.0	15.3 (7.5-23.4)	11.4 (5.1-19.0)	6.3 (2.1-11.6)	1.6 (0.2-3.5)
mtDNA (no data)	BEAST (strict clock)	n/a	16.6±0.1	16.6 (16.4-16.8)	10.1 (3.3-16.6)	4.2 (0.0-11.0)	4.2 (0.0-11.0)
nuclear	BEAST (strict clock)	-15611.76	16.6±0.1	16.6 (16.4-16.8)	8.6 (6.1-11.0)	3.5 (2.0-5.2)	n/a
nuclear (no data)	BEAST (strict clock)	n/a	16.6±0.1	16.6 (16.4-16.8)	9.0 (2.1-16.6)	3.9 (0.0-10.6)	n/a
nuclear	BEST	-15949.8	16.6	16.6 (13.9-19.2)	12.2 (9.9-14.4)	6.2 (3.0-9.5)	n/a

Table 3. Likelihood and coalescent branch lengths for the species tree analyses. Values are means and 95% highest posterior density intervals.

mtDNA	clade B1 alleles	harmonic mean LnL	mean coalescent branch lengths (95% HPD)				
			C. intermedia	*C. orinocensis*	ancestor of *intermedia+ orinocensis*	*C. oc. monoculus*	ancestor of *intermedia+ monoculus*
no	no	-15949.8	1.017 (0.492 - 1.826)	0.726 (0.356 - 1.253)	2.519 (1.169 - 5.049)	0.725 (0.407 - 1.144)	2.398 (1.61 - 3.447)
no	yes	-16042.912	1.041 (0.558 - 2.066)	0.417 (0.214 - 0.688)*	2.309 (1.339 - 3.809)	0.877 (0.516 - 1.452)	2.376 (1.597 - 3.385)
yes	no	-17659.3	1.181 (0.599 - 2.098)	0.884 (0.476 - 1.486)	3.287 (1.373 - 8.082)	0.807 (0.48 - 1.229)	2.657 (1.771 - 3.948)
yes	yes	-17759.629	1.384 (0.685 - 2.474)	0.439 (0.244 - 0.68)*	1.265 (0.708 - 1.919)**	0.645 (0.354 - 1.103)	1.536 (1.079 - 2.06)**

Figure Legends

Figure 1. Models of gene trees (bold lines) in species trees (solid thin lines) for hypotheses of A) deep coalescence or incomplete lineages sorting and B) ancient introgression. In each case, the divergence time of clade B1 alleles in *C. orinocensis* and *C. oc. monoculus* is indicated (T_{B1}), along with the speciation age of the clade B species (τ_{cladeB}).

Figure 2. Results from previous population-level analyses of *Cichla orinocensis*, modified from Chapter 2. A) Distribution of *C. orinocensis* with sampling localities classified by the mtDNA clade found in each. B) MtDNA genealogy of the genus *Cichla*, with the two major clades of *C. orinocensis* identified. This phylogram was inferred using a maximum likelihood (ML) search of mtDNA control region (CR) haplotypes concatenated with ATPase 8,6 sequences. Values associated with key branches are ML bootstrap percentages. The width of collapsed clades is not indicative of internal diversity. Terminals are CR haplotypes from: clade B1: 43 individuals; clade B2: 192 individuals; 9 recent hybrids grouped within *C. ocellaris s.l.* and are not shown. C) Diagram of the optimal clustering (K=2) of *C. orinocensis* individuals based on analysis of 12 microsatellite loci with the program STRUCTURE (Pritchard et al. 2000). External lines denote sampling localities, and symbols mark the mtDNA clade found in each.

Figure 3. Time-calibrated trees for the mtDNA (ATPase) and organismal phylogeny. Node bars represent the 95% highest posterior density intervals for divergence time.

Figure 4. Histogram of the results from the filtering of topologies of the genealogies simulated with MESQUITE under a multispecies coalescent model.

Figure 1

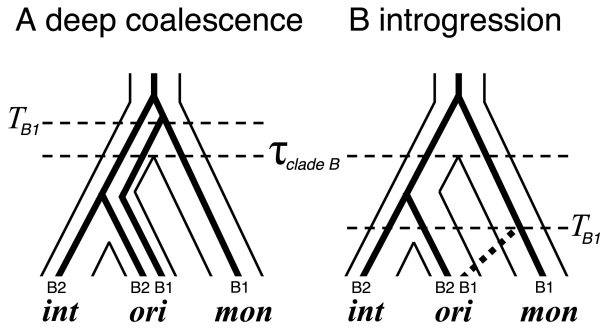

A deep coalescence
B introgression
T_{B1}
$\tau_{clade\ B}$
T_{B1}
B2
B2 B1
B1
int
ori
mon
B2
B2 B1
B1
int
ori
mon

Figure 2

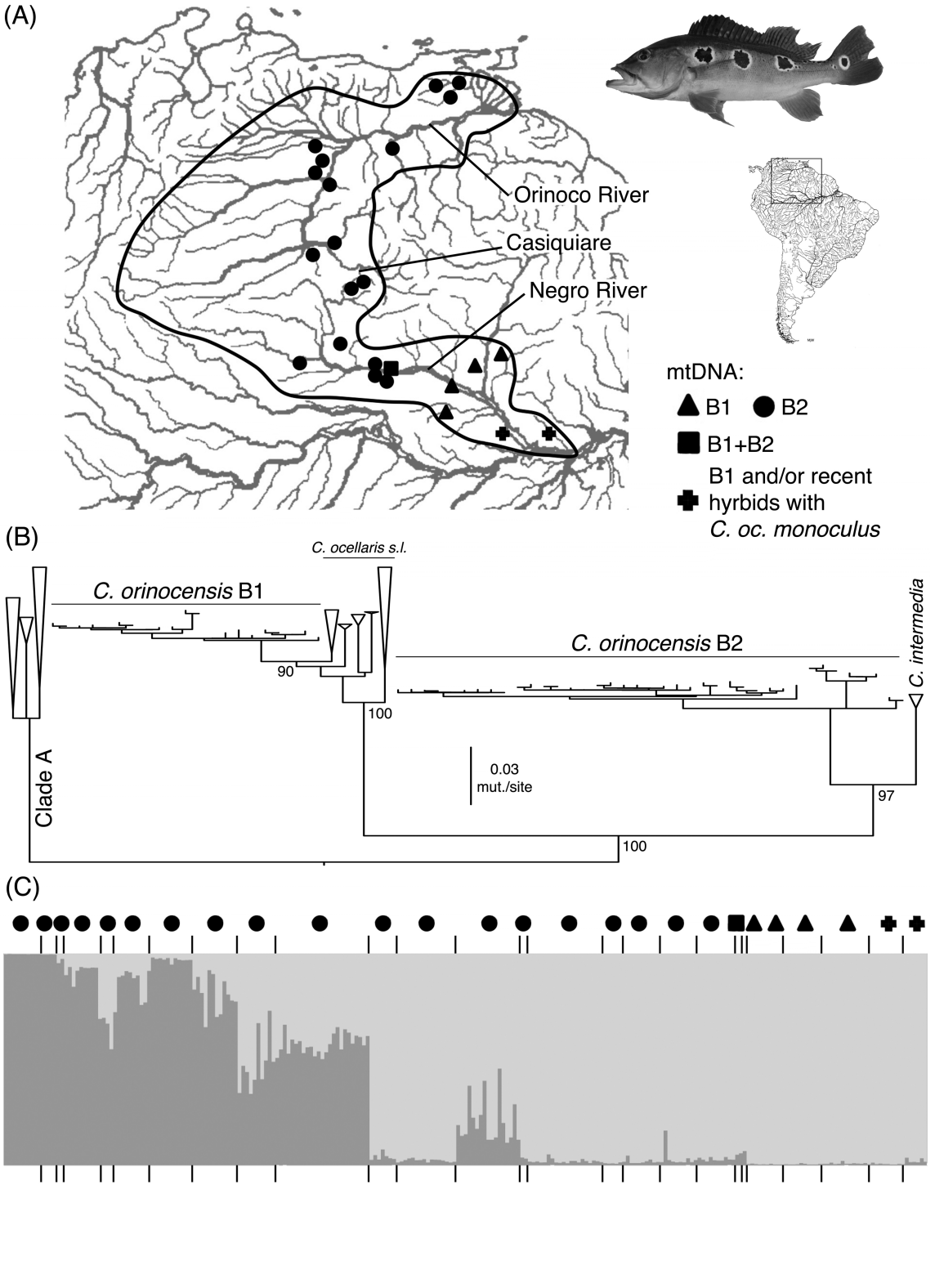

(A)
Orinoco River
Casiquiare
Negro River
mtDNA:
B1
B2
B1+B2
B1 and/or recent hyrbids with C. oc. monoculus
(B)
C. ocellaris s.l.
C. orinocensis B1
C. orinocensis B2
C. intermedia
Clade A
90
100
97
100
0.03 mut./site
(C)

Figure 3

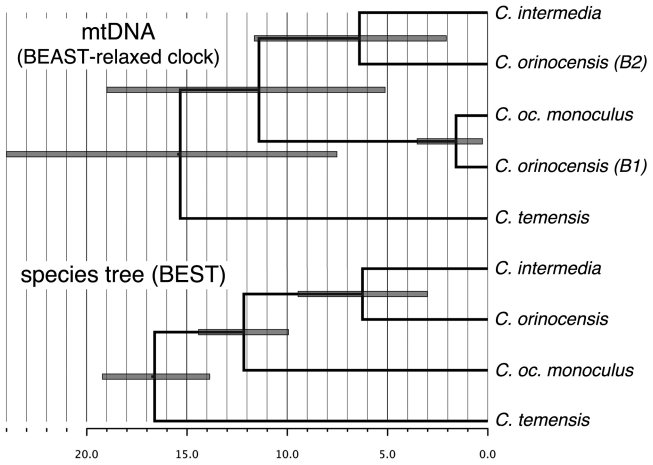
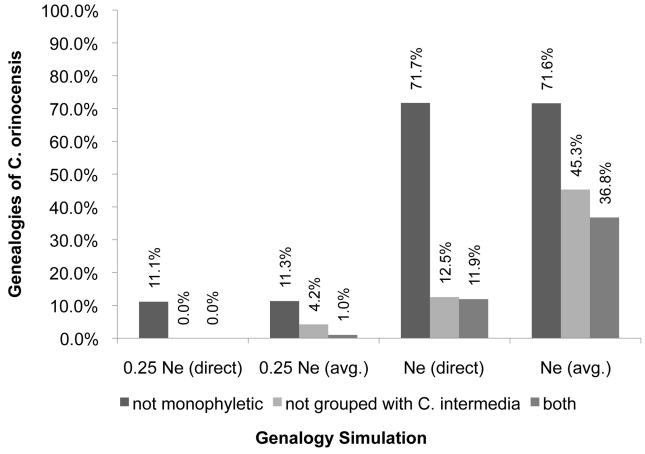


Figure 4



Literature Cited

Akaike, H. 1974. A new look at the statistical model identification. IEEE Transactions of Automatic Control 19:716-723.

Albert, J., and W. G. R. Crampton. 2005. Diversity and phylogeny of Neotropical electric fishes (Gymnotiformes). Pp. Chapter 13 *in* T. H. Bullock, C. D. Hopkins, A. N. Popper, and R. R. Fay, eds. Electroreception. Springer.

Anderson, E., and G. L. Stebbins. 1954. Hybridization as an evolutionary stimulus. Evolution 8:378-388.

Ane, C., B. Larget, D. A. Baum, S. D. Smith, and A. Rokas. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution 24:412-426.

Apagow, P. M., O. R. P. Bininda-Emonds, K. A. Crandall, J. L. Gittleman, G. M. Mace, J. C. Marshall, and A. Purvis. 2004. The impact of species concept on biodiversity studies. Quarterly Review of Biology 79:161-179.

Arnold, M. L. 1997. Natural Hybridization and Evolution. Oxford University Press, Oxford, U.K.

Arnold, M. L., and N. H. Martin. 2009. Adaptation by introgression. Journal of BIology 8:82.

Avise, J. C., J. Arnold, R. M. Ball, E. Bermingham, T. Lamb, J. E. Neigel, C. A. Reeb, and N. C. Saunders. 1987. Intraspecific phylogeography: The mitochondrial DNA bridge between population genetics and systematics. Annual Review of Ecology and Systematics 18:489-522.

Avise, J. C., and R. M. Ball. 1990. Principles of genealogical concordance in species concepts and biological taxonomy. Pp. 45-67 *in* D. Futuyma, and J. Antonovics, eds. Oxf Surv Evol Biol. Oxford University Press, Oxford.

Bachtrog, D., K. Thornton, A. Clark, and P. Andolfatto. 2006. Extensive introgression of mitochondrial DNA relative to nuclear genes in the *Drosophila yakuba* species group. Evolution 60:292-302.

Baker, R., and R. DeSalle. 1997. Multiple sources of character information and the phylogeny of Hawaiian drosopholids. Systematic Biology 46:654-673.

Ballard, J. W. O., and M. C. Whitlock. 2004. The incomplete natural history of mitochondria. Molecular Ecology 13.

Baum, D. A., and K. L. Shaw. 1995. Genealogical perspectives on the species problem *in* P. C. Hoch, A. G. Stevenson, and B. A. Schaal, eds. Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St. Louis.

Bermerguy, R. L., and J. B. Sena Costa. 1991. Considerações sobre a evolução do sistema de drenagen da Amazônia e sua relação com o arcabouço tectônico-estrutual. Museo Paraense Emilio Gouldi, Serio Ciencias da Terra 3:75-97.

Bermingham, E., and A. P. Martin. 1998. Comparative mtDNA phylogeography of Neotropical freshwater fishes: testing shared history to infer the evolutionary landscape of lower Central America. Molecular Ecology 7:499-517.

Brinn, M. N. A., J. I. R. Porto, and E. Feldberg. 2004. Karyological evidence for interspecific hybridization between *Cichla monoculus* and *C. temensis* (Perciformes, Cichlidae) in the Amazon. Hereditas (Lund) 141:252-257.

Brown, J. M., and A. R. Lemmon. 2007. The importance of data partitioning and the utility of bayes factors in Bayesian phylogenetics. Syst Biol 56:643-655.
Brown, R. P., and Z. Yang. 2010. Bayesian dating of shallow phylogenies with a relaxed clock. Systematic Biology 59:119-131.
Brumfield, R. T., L. Liu, D. E. Lum, and S. V. Edwards. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, Manacus) from multilocus sequence data. Syst Biol 57:719-731.
Buckley, T. R., M. Cordeiro, D. C. Marshall, and C. Simon. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (Maoricicada Dugdale). Syst Biol 55:411-425.
Buckley, T. R., C. Simon, and G. K. Chambers. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50:67-86.
Buerkle, C. A., R. J. Morris, M. A. Asmussen, and L. H. Rieseberg. 2000. The likelihood of homoploid hybrid speciation. Heredity 84 (Pt 4):441-451.
Bull, J. J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddel. 1993. Partitioning and combining data in phylogenetic analysis. Systematic Biology 42:384-397.
Bull, V., M. Beltran, C. D. Jiggins, W. O. McMillan, E. Bermingham, and J. Mallet. 2006. Polyphyly and gene flow between non-sibling Heliconius species. BMC Biol 4:11.
Burridge, C. P., D. Craw, D. C. Jack, T. M. King, and J. M. Waters. 2008. Does fish ecology predict dispersal across a river drainage divide? Evolution 62:1484-1499.
Burridge, C. P., D. Craw, and J. M. Waters. 2006. River capture, range expansion, and cladogenesis: the genetic signature of freshwater vicariance. Evolution 60:1038-1049.
Burridge, C. P., D. Craw, and J. M. Waters. 2007. An empirical test of freshwater vicariance via river capture. Molecular Ecology 16:1883-1895.
Carstens, B. C., and T. A. Dewey. 2010. Species delimitation using a combined coalescent and information-theoretic approach: an example from North American *Myotis* bats. Systematic Biology 59:400-414.
Chan, K. M. A., and S. A. Levin. 2005. Leaky prezygotic isolation and porous genomes: rapid introgression of maternally inherited DNA. Evolution 59:720-729.
Chernoff, B., A. Machado-Allison, and W. G. Saul. 1991. Morphology, variation and biogeography of *Leporinus brunneus* (Pices: Characiformes: Anostomidae). Ichthyological Exploration of Freshwaters 1:295-306.
Clement, M., D. Posada, and K. A. Crandall. 2000. TCS: a computer program to estimate gene genealogies. Molecular Ecology 9:1657-1659.
Collar, D. C., T. J. Near, and P. C. Wainwright. 2005. Comparative analysis of morphological diversity: does disparity accumulate at the same rate in two lineages of centrarchid fishes? Evolution 59:1783-1794.
Collar, D. C., C. O'Meara B, P. C. Wainwright, and T. J. Near. 2009. Piscivory limits diversification of feeding morphology in centrarchid fishes. Evolution 63:1557-1573.

Cooper, J. D., R. Vitalis, P. M. Waser, D. Gopurenko, E. C. Hellgren, T. M. Gabor, and J. A. DeWoody. 2010. Quantifying male-biased dispersal among social groups in the collared peccary (*Pecari tajacu*) using analyses based on mtDNA variation. Heredity 104:79-87.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer, Sunderland, MA.

Coyne, J. A., H. A. Orr, and D. J. Futuyma. 1988. Do we need a new species concept? Systematic Zoology 37:190-200.

Cracraft, J. 1989. Speciation and its ontology: The empirical consequenses of alternative species concepts for understanding patterns and processes of differentiation. Pp. 28-59 *in* D. Otte, and J. A. Endler, eds. Speciation and its consequences. Sinauer, Sunderland, Massachusetts.

de Queiroz, K. 1998. The general lineage concept of species, species criteria, and the process of speciation: a conceptual unification and terminological recommendations *in* D. J. Howard, and S. H. Berlocher, eds. Endless Forms: Species and Speciation. Oxford University Press, Oxford, England.

Degnan, J. H., and M. S. Rosenberg. 2009. Gene tree discordance, phylogenetic inference, and the multispecies coalescent. Trends in Ecology & Evolution 24:332-340.

Degnan, J. H., and N. A. Rosenberg. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet 2:e68.

Degnan, J. H., and L. A. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24-37.

Di Candia, M. R., and E. Routman. 2007. Cytonuclear discordance across a leopard frog contact zone. Molecular Phylogenetics and Evolution 45:564-575.

Dobzhansky, T. 1937. Genetics and the Origin of Species. Columbia University Press, New York, NY.

Donnelly, M. J., J. Pinto, R. Girod, N. J. Besansky, and T. Lehmann. 2004. Revisiting the role of introgression vs shared ancestral polymorphisms as key processes shaping genetic diversity in the recently separated sibling species of the *Anopheles gambiae* complex. Heredity 92:61-68.

Dowling, D. K., U. Friberg, and J. Lindell. 2008. Evolutionary implications of non-neutral mitochondrial genetic variation. Trends in Ecology & Evolution 23:546-554.

Dowling, T. E., and C. L. Secor. 1997. The role of hybridization and introgression in the diversification of animals. Annual Review of Ecology and Systematics 28:593-619.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond, A. J., S. Y. W. Ho, N. Rawlence, and A. Rambaut. 2007. A Rough Guide to BEAST 1.4, Auckland, New Zealand.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214.

Eckert, A. J., and B. C. Carstens. 2008. Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. Molecular Phylogenetics and Evolution 49:832-842.

Edwards, C. E., D. E. Soltis, and P. S. Soltis. 2008. Using patterns of genetic structure based on microsatellite loci to test hypotheses of current hybridization, ancient hybridization, and incomplete lineage sorting in Conradina (Laminaceae). Molecular Ecology 17:5157-5174.
Edwards, S. V. 2008. Is a new and general theory of molecular systematics emerging? Evolution 63:1-19.
Edwards, S. V., and P. Beerli. 2000. Gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54:1839-1854.
Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proc Natl Acad Sci U S A 104:5936-5941.
Ellegren, H. 2004. Microsatellites: simple sequences with complex evolution. Nature Reviews Genetics 5:435-445.
Ellegren, H., C. R. Primmer, and B. C. Sheldon. 1995. Microsatellite 'evolution': Directionality or bias? Nature Genetics 11:360-362.
Evanno, G., S. Regnaut, and J. Goudet. 2005. Detecting the number of clusters of individuals using the software Structure: a simulation study. Molecular Ecology 14:2611-2620.
Farias, I. P., G. Orti, I. Sampaio, H. Schneider, and A. Meyer. 1999. Mitochondrial DNA phylogeny of the family Cichlidae: monophyly and fast molecular evolution of the Neotropical assemblage. Journal of Molecular Evolution 48:703-711.
Felsenstein, J. 1985a. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783-791.
Felsenstein, J. 1985b. Phylogenies and the comparative method. American Naturalist 125:1-15.
Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.
Freeman, B., L. G. Nico, M. Osentoski, H. L. Jelks, and T. M. Collins. 2007. Molecular systematics of Serrasalmidae: Deciphering the identities of piranha species and unraveling their evolutionary histories. Zootaxa:1-38.
Funk, D. J., and K. E. Omland. 2003. SPECIES-LEVEL PARAPHYLY AND POLYPHYLY: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397-423.
Galtier, N., B. Nabholz, S. Glemin, and G. D. D. Hurst. 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Molecular Ecology 18:4541-4550.
Garrick, R. C., P. Sunnucks, and R. J. Dyer. 2010. Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. BMC Evolutionary Biology 10:118.
Gerber, A., R. Loggins, S. Kumar, and T. E. Dowling. 2001. Does nonneutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes. Annual review of Genetics 35:539-566.
Gery, J. 1969. The Fresh-water Fishes of South America *in* E. J. Fittkau, J. Illies, H. Klinge, G. H. Schwabe, and H. Sioli, eds. Biogeography and Ecology in South America. Dr. W. Junk, The Hague.
Goldstein, D. B., and D. D. Pollock. 1997. Launching microsatellites: a review of mutation processes and methods of phylogenetic inference. Journal of Heredity 88:335-342.

Goldstein, D. B., and C. Schlötterer, eds. 1999. Microsatellites: Evolution and Applications. Oxford University Press, Oxford.

Good, J., S. Hird, N. Reid, J. R. Demboski, S. J. Steppan, T. R. Martin-Nims, and J. Sullivan. 2008. Ancient hybridization and mitochondrial capture between two species of chipmunks. Molecular Ecology 17:1313-1327.

Grant, P., B. R. Grant, and K. Petren. 2005. Hybridization in the recent past. American Naturalist 166:56-67.

Graur, D., and W. Martin. 2004. Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. Trends in Genetics 20:80-86.

Hansson, B., D. Hasselquist, M. Tarka, P. Zehtindjiev, and S. Bensch. 2008. Postglacial colonisation patterns and the role of isolation and expansion in driving diversification in a passerine bird. PLoS ONE 3:e2794.

Harrison, R. G., and A. Hastings. 1996. Genetic and evolutionary consequences of metapopulation structure. Trends in Ecology & Evolution 11:180-183.

Harrison, R. G., and D. M. Rand. 1989. Mosaic hybrid zones and the nature of species boundaries. Pp. 111-133 *in* D. Otte, and J. A. Endler, eds. Speciation and Its Consequences. Sinauer Associates, Sunderland, MA.

Harvey, P. H., and M. D. Pagel. 1991. The Comparative Method in Evolutionary Biology. Oxford University Press, New York.

Hausdorf, B., and C. Hennig. 2010. Species delimitation using dominant and codominant multilocus markers. Systematic Biology 59:491-503.

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species tree from multilocus data. Molecular Biology and Evolution 27:570-580.

Hey, J., and R. Nielsen. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the Natinal Academy of Sciences USA 104:2785-2790.

Hey, J., R. S. Waples, M. L. Arnold, R. K. Butlin, and R. G. Harrison. 2003. Understanding and confronting species uncertainty in biology and conservation. Trends in Ecology & Evolution 18:597-603.

Hey, J., Y. J. Won, A. Sivasundar, R. Nielsen, and J. A. Markert. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated populations. Molecular Ecology 13:909-919.

Hickerson, M. J., C. P. Meyer, and C. Moritz. 2006. DNA barcoding will often fail to discover new animal species over broad parameter space. Syst Biol 55:729-739.

Ho, S. Y. W., and M. J. Phillips. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Systematic Biology 58:367-380.

Hoeinghaus, D. J., C. A. Layman, D. A. Arrington, and K. O. Winemiller. 2003. Movement of *Cichla* species (Cichlidae) in a Venezuelan floodplain river. Neotropical Ichthyology 1:121-126.

Hoorn, C. 1994. An Environmental Reconstruction of the Palaeo-Amazon River System (Middle-Late Miocene, Nw Amazonia). Palaeogeography Palaeoclimatology Palaeoecology 112:187-238.

Hoorn, C., J. Guerrero, G. A. Sarmiento, and M. A. Lorente. 1995. Andean Tectonics as a Cause for Changing Drainage Patterns in Miocene Northern South-America. Geology 23:237-240.

Hoorn, C., F. Wesselingh, H. ter Steege, M. A. Bermudez, A. Mora, J. Sevink, I. Sanmartín, A. Sanchez-Meseguer, C. L. Anderson, J. P. Figueiredo, C. Jaramillo, D. Riff, F. R. Negri, H. Hooghiemstra, J. G. Lundberg, T. Stadler, T. Säkinen, and A. Antonelli. 2010. Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. Science 330:927-931.

Huang, H., and L. L. Knowles. 2009. What is the danger of the anomaly zone for empirical phylogenetics? Systematic Biology 58:527-536.

Hubert, N., and J. F. Renno. 2006. Historical biogeography of South American freshwater fishes. Journal of Biogeography 33:1414-1436.

Hubisz, M. J., D. Falush, M. Stephens, and J. K. Pritchard. 2009. Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9:1322-1332.

Hudson, R. R., and N. I. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147-164.

Huelsenbeck, J. P., and P. Andolfatto. 2007. Inference of population structure under a Dirichlet process model. Genetics 175:1787-1802.

Huelsenbeck, J. P., J. P. Bollback, and A. M. Levine. 2002. Inferring the root of a phylogenetic tree. Systematic Biology 51:32-43.

Huelsenbeck, J. P., and J. J. Bull. 1996. A likelihood ratio test for detection of conflicting phylogenetic signal. Syst. Biol. 45:92-98.

Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science (Washington D C) 276:227-232.

Hull, D. L. 1977. The ontological status of species as evolutionary units. Pp. 91-102 *in* R. Butts, and J. Hintikka, eds. Foundational Problem in the Special Sciences. D. Reidel Publishing Company, Dordrecht, Holland.

Isaac, N. J., J. Mallet, and G. M. Mace. 2004. Taxonomic inflation: its influence on macroecology and conservation. Trends in Ecology & Evolution 19:464-469.

Jakobsson, M., and N. A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801-1806.

Jepsen, D. B., K. O. Winemiller, and D. C. Taphorn. 1997. Temporal patterns of resource partitioning among Cichla species in a Venezuelan blackwater river. Journal of Fish Biology 51:1085-1108.

Jobb, G., A. von Haeseler, and K. Strimmer. 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. Bmc Evolutionary Biology 4:-.

Karl, S. A., and B. Bowen. 1999. Evolutionary significant units versus geopolitical taxonomy: molecular systematics of an endangered sea turle (genus *Chelonia*). Conservation Biology 13:990-999.

Kass, R. E., and A. E. Raftery. 1995. Bayes Factors. Journal of the American Statistical Association 90:773-795.

Katoh, K., K. Kuma, H. Toh, and T. Miyata. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acid Research 33:511-518.

Kelchner, S. A., and M. A. Thomas. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22:87-94.

Kimura, M., and J. Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics 49:725-738.
Kingman, J. F. C. 1982. The coalescent. Stochastic Processes and Their Applications 13:235-248.
Kishino, H., and M. Hasegawa. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. Journal of Molecular Evolution 29:170-179.
Knowles, L. L., and B. C. Carstens. 2007. Delimiting species without monophyletic gene trees. Syst Biol 56:887-895.
Knowles, L. L., and D. R. Maddison. 2002. Statistcal phylogeography. Molecular Ecology 11:2623-2635.
Koonin, E. V. 2005. Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics 39:309-338.
Kubatko, L. S., B. C. Carstens, and L. L. Knowles. 2009. STEM: Species Tree Estimation using Maximum likelihood for gene trees under coalescence. Bioinformatics doi: 10.1093/bioinformatics/btp079.
Kubatko, L. S., and J. H. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56:17-24.
Kuhner, M. K. 2006. LAMARC 2.0: maximum likelihood and Bayesian estimation of population parameters. Bioinformatics 22:768-770.
Kuhner, M. K., J. Yamato, and J. Felsenstein. 1995. Estimating effective population size and neutral mutation rate from sequence data using Metropolis-Hastings sampling. Genetics 140:1421-1430.
Kullander, S. O., and E. J. G. Ferreira. 2006. A review of the South American cichlid genus Cichla, with descriptions of nine new species (Teleostei : Cichlidae). Ichthyological Exploration of Freshwaters 17:289-398.
Lacy, R. C. 1987. Loss of genetic diversity from manages populations: interaction effects of drift, mutation, immigration, selection, and population subdivision. Conservation Biology 1:143-158.
Layman, C. A., and K. O. Winemiller. 2004. Size-based responses of prey to piscivore exclusion in a species-rich neotropical river. Ecology 85:1311-1320.
Lee, J. Y., and S. V. Edwards. 2008. Divergence across Australia's Carpenterian barrier: statistical phylogeography of the red-backed fairy wren (Malurus melanocephalus). Evolution 62:3117-3134.
Lewontin, R. C., and L. C. Birch. 1966. Hybridization as a source of variation for adaptation to new environments. Evolution 20:315-336.
Li, C., G. Lu, and G. Ortí. 2008. Optimal data partitioning and a test case for ray-finned fishes (Actinopterygii) based on ten nuclear loci. Systematic Biology 57:519-539.
Li, C., J.-J. M. Riethoven, and L. Ma. 2010. Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. BMC Evolutionary Biology 10:90.
Li, Y.-C., A. B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Molecular Ecology 11:2453-2465.
Linnen, C. R., and B. D. Farrell. 2007. Mitonuclear discordance is caused by rampant mitochondrial introgression in *Neodiprion* (Hymenoptera: Diprionidae) sawflies. Evolution 61:1417-1438.

Liu, L., and D. K. Pearl. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Systematic Biology 56:504-514.
Liu, L., D. K. Pearl, R. T. Brumfield, and S. V. Edwards. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080-2091.
López-Fernández, H., R. L. Honeycutt, and K. O. Winemiller. 2005. Molecular phylogeny and evidence for an adaptive radiation of geophagine cichlids from South America (Perciformes : Labroidei). Molecular Phylogenetics and Evolution 34:227-244.
López-Fernández, H., K. O. Winemiller, and R. L. Honeycutt. 2010. Multilocus phylogeny and rapid radiations in Neotrpical cichlid fishes (Perciformes: Cichlidae: Cichlinae). Molecular Phylogenetics and Evolution 55:1070-1086.
Lopez-Rojas, H., A. Machado-Allison, and F. Mago-Leccia. 1978. Review of Ecological Studies in Tropical Fish Communities by R.H. Lowe-McConnell. Copeia 1988:503-505.
Lovejoy, N. R., and M. L. G. de Araújo. 2000. Molecular systematics, biogeography and population structure of Neotropical freshwater needlefishes of the genus *Potamorrhaphis*. Molecular Ecology 9:259-268.
Lundberg, J. G. 1997. Fishes of the La Venta Fauna: additional taxa, biotic and paleoenvironmental implications. Pp. 67-91 *in* R. H. M. R. F. Kay, R. L. Cifelli and J. J. Flynn, ed. Vertebrate paleontology in the Neotropics: The Miocene fauna of La Venta Colombia. Smithsonian Institution Press, Washington, D.C.
Lundberg, J. G. 1998. The Temporal Context for the Diversification of Neotropical Fishes. Pp. 49-68 *in* L. R. Malabarba, R. E. Reis, R. P. Vari, C. A. S. Lucena, and Z. M. S. Lucena, eds. Phylogeny and Classification of Neotropical Fishes.
Lundberg, J. G., L. G. Marshall, J. Guerrero, B. Horton, M. C. S. L. Malabarba, and F. Wesselingh. 1998. The Stage for Neotropical Fish Diversification: A History of Tropical South American Rivers. Pp. 13-48 *in* L. R. Malabarba, R. E. Reis, R. P. Vari, C. A. S. Lucena, and Z. M. S. Lucena, eds. Phylogeny and Classification of Neotropical Fishes.
Maddison, W., and D. R. Maddison. 2008. Mesquite: a modular system for evolutionary analysis. Version 2.5.
Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46:523-536.
Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21-30.
Maddison, W. P., and D. R. Maddison. 1992. MacClade: Analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Massachusetts.
Mago-Leccia, F. 1971. La ictiofauna del Casiquiare. Revista Defense de la Naturaleza, Caracas 1:5-10.
Mäkinen, H. S., and J. Merilä. 2008. Mitochondrial DNA phylogeography of the three-spined stickleback (*Gasterosteus aculeatus*) in Europe—Evidence for multiple glacial refugia. Molecular Phylogenetics and Evolution 46:167-182.
Mallet, J. 2005. Hybridization as an invasion of the genome. Trends Ecol Evol 20:229-237.
Mallet, J. 2007a. Hybrid speciation. Nature 446:279-283.

Mallet, J. 2007b. Subspecies, semispecies, superspecies. Pp. 1-5 *in* S. A. Levin, ed. Encyclopedia of Biodiversity. Elsevier, Oxford.
Mallet, J., M. Beltran, W. Neukirchen, and M. Linares. 2007. Natural hybridization in heliconiine butterflies: the species boundary as a continuum. BMC Evol Biol 7:28.
Mayr, E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Harvard University Press, Cambridge, MA.
Mayr, E. 1963. Animal Species and Evolution. Harvard University Press, Cambridge, MA.
Meyer, A. 1993. Evolution of mitochondrial DNA of fishes. Pp. 1-38 *in* P. W. Hochachka, and P. Mommsen, eds. Molecular Biology Frontiers, Biochemistry and Molecular Biology of Fishes. Elsevier Press., Amsterdam.
Mishler, B. D., and M. J. Donoghue. 1982. Species concepts: a case for pluralism. Systematic Zoology 31:491-503.
Morando, M., L. J. Avila, J. Baker, and J. W. Sites, Jr. 2004. Phylogeny and phylogeography of the Liolaemus darwinii complex (Squamata: Liolaemidae): evidence for introgression and incomplete lineage sorting. Evolution 58:842-861.
Morando, M., L. J. Avila, and J. W. Sites, Jr. 2003. Sampling strategies for delimiting species: genes, individuals, and populations in the Liolaemus elongatus-kriegi complex (Squamata: Liolaemidae) in Andean-Patagonian South America. Syst Biol 52:159-185.
Moritz, C. 1994. Defining 'evolutionarily siginificant units' for conservation. Trends in Ecology & Evolution 9:373-375.
Nei, M., and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA 76:5269-5273.
Nichols, R. 2001. Gene trees and species trees are not the same. Trends in Ecology & Evolution 16:358-364.
Nielsen, R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor Popul Biol 53:143-151.
Nylander, J. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst Biol 53:47-67.
Nylander, J. A., J. C. Wilgenbusch, D. L. Warren, and D. L. Swofford. 2008. AWTY (Are We There Yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24:581-583.
Nylander, J. A. A. 2004. MrModeltest v2. Program distributed by the author, Evolutionary Biology Centre, Uppsala University.
O'Meara, B. C. 2010. New heuristic methods for joint species delimitation and species tree inference. Systematic Biology 59:59-73.
Oliveira, A. V., A. J. Prioli, S. M. A. P. Prioli, T. S. Bignotto, H. F. Julio Jr., H. Carrer, C. S. Agostinho, and L. M. Prioli. 2006. Genetic diversity of invasive and native Cichla (Pices: Perciformes) populations in Brazil with evidence of interspecific hybridization. Journal of Fish Biology 69:260-277.
Padial, J. M., and I. de la Riva. 2006. Taxonomic inflation and the stability of species lists: the perils of ostrich's behavior. Systematic Biology 55:859-867.

Page, R. D. M., and J. C. Cotton. 2000. GeneTree: a tool for exploring gene family evolution. Pp. 525-536 *in* D. Sankoff, and J. Nadeau, eds. Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families. Kluwer Academic Publishers, Dordrecht.

Pagel, M., and A. Meade. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571-581.

Palumbi, S. R., F. Cipriano, and M. P. Hare. 2001. Predicting nuclear gene coalescence from mitochondrial data: the three times rule. Evolution 55.

Pamilo, P., and M. Nei. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5:568-583.

Peters, J. L., W. Gretes, and K. E. Omland. 2005. Late Pleistocene divergence between eastern and western populations of wood ducks (*Aix sponsa*) inferred by the 'isolation with migration' coalescent method. Molecular Ecology 14:3407-3418.

Peters, J. L., Y. N. Zhuravlev, I. Fefelov, A. Logie, and K. E. Omland. 2007. Nuclear loci and coalescent methods support ancient hybridization as cause of mitochondrial paraphyly between gadwall and falcated duck (*Anas* spp.). Evolution 61:1992-2006.

Pons, J., T. G. Barraclough, J. Gomez-Zurita, A. Cardoso, D. P. Duran, S. Hazell, S. Kamoun, W. D. Sumlin, and A. P. Vogler. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst Biol 55:595-609.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics 14:817-818.

Pritchard, J. K., M. Stephens, and P. J. Donnelly. 2000. Inference of populations structure using multilocus genotype data. Genetics 155:945-959.

Raftery, A. E. 1996. Hypothesis testing and model selection. Pp. 163-188 *in* W. R. Gilks, D. J. Spiegelhalter, and S. Richardson, eds. Markov Chain Monte Carlo in Practice. Chapman and Hall, London.

Rambaut, A., and A. J. Drummond. 2007. Tracer v1.4. Available from http://beast.bio.ed.ac.uk/Tracer.

Randler, C. 2001. Avian hybridization, mixed pairing and female choice. Anim Behav 63:103-119.

Rannala, B. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Syst. Biol. 51:754-760.

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

Reed, R. D., and F. A. Sperling. 1999. Interaction of process partitions in phylogenetic analysis: an example from the swallowtail butterfly genus Papilio. Mol. Biol. Evol. 16:286-297.

Reis, R. E., S. O. Kullander, and C. J. Ferraris, eds. 2003. Checklist of the Freshwater Fishes of South and Central America. EDIPUCRS, Porto Alegre, Brazil.

Renno, J. F., N. Hubert, J. P. Torrico, F. Duponchelle, J. N. Rodriguez, C. G. Davila, S. C. Willis, and E. Desmarais. 2006. Phylogeography of Cichla (Cichlidae) in the upper Madera basin (Bolivian Amazon). Molecular Phylogenetics and Evolution 41:503-510.

Rice, A. H. 1921. The Rio Negro, the Casiquiare Canal, and the Upper Orinoco, September 1919-April 1920. The Geographic Journal 58:321-343.
Rieseberg, L., J. Whitton, and C. R. Linder. 1996. Molecular marker incongruence in plant hybrid zones and phylogenetic trees. Acta Botanica Neerlandica 45:243-262.
Roberts, T. R., E. J. Sargis, and L. E. Olson. 2009. Networks, trees, and treeshrews: Assessing support and identifying conflict with multiple loci and a problematic root. Systematic Biology 58:257-270.
Rokas, A., and S. B. Carroll. 2005. More sequence or more taxa? The relative contribution of taxon number and sequences data size to phylogenetic accuracy. Molecular Biology and Evolution 22:1337-1344.
Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798-804.
Ronquist, F. 1997. Dispersal-vicariance analysis: a new approach to the quantification of historical biogeography. Systematic Biology 46:195-203.
Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.
Rosenberg, N. A., and M. Nordborg. 2002. Genealogical trees, coalescent theory, and the analysis of genetic polymorphisms. Nature Reviews Genetics 3:380-390.
Rozas, J., J. C. Sánchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496-2497.
Rubinsztein, D. C., B. Amos, and G. Cooper. 1999. Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages. Philos Trans R Soc Lond B Biol Sci 354:1095-1099.
Rubinsztein, D. C., B. Amos, J. Leggo, S. Goodburn, S. Jain, S.-H. Li, R. L. Margolis, C. A. Ross, and M. A. Furguson-Smith. 1995. Microsatellite evolution- evidence for directionality and variation in rate between species. Nature Genetics 10:337-343.
Shaw, K. L. 2001. The genealogical view of speciation. Journal of Evolutionary Biology 14:880-882.
Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.
Sides, J., and C. Lydeard. 1999. Phylogenetic utility of the tyrosine kinase gene X-src for assessing relationships among representative cichlid fishes. Molecular Phylogenetics and Evolution 14:51-74.
Simpson, G. G. 1961. Principles of Animal Taxonomy. Columbia University Press, New York, NY.
Sioli, H. 1984. The Amazon and its main affluents: hydrography, morphology of the river courses, and river types. Pp. 127-163 *in* H. Sioli, ed. The Amazon: Limnology and Landscape Ecology of a Mighty Tropical River and its Basin. Dr. W. Junk Publishers, The Hague.
Sites, J. L., and K. Crandall. 1997. Testing species boundaries in biodiversity studies. Conservation Biology 11:1289-1297.
Sites, J. W., Jr., and J. C. Marshall. 2003. Delimiting species: a Renaissance issue in systematic biology. Trends in Ecology & Evolution 18:462-470.

Slowinski, J. B., and R. D. Page. 1999. How should species phylogenies be inferred from sequence data? Syst Biol 48:814-825.
Stelkens, R., and O. Seehausen. 2009. Genetic distance between species predicts novel trait expression in their hybrids. Evolution 63:884-897.
Stephens, M., N. J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. American Journal of Human Genetics 68:978-989.
Stern, K. 1970. Der Casiquiare-Kanal, einst und jetzt. Amazoniana 2:401-416.
Strasburg, J. L., and L. H. Rieseberg. 2008. Molecular demographic history of the annual sunflowers Helianthus annuus and H. petiolaris--large effective population sizes and rates of long-term gene flow. Evolution 62:1936-1950.
Sturmbauer, C., S. Baric, W. Salzburger, L. Ruber, and E. Verheyen. 2001. Lake level fluctuations synchronize genetic divergences of cichlid fishes in African lakes. Mol. Biol. Evol. 18:144-154.
Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445-466.
Swofford, D. L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.
Takahata, N. 1989. Gene genealogy in three related populations: Consistency probability between gene and population trees. Genetics 122:957-966.
Templeton, A. 2001. Using phylogeographic analyses of gene trees to test species status and processes. Molecular Ecology 10:779-791.
Templeton, A. R. 1989. The meaning of species and speciation: A genetic perspective. Pp. 3-27 *in* D. Otte, and J. A. Endler, eds. Speciation and its consequences. Sinauer, Sunderland, Massachusetts.
Thacker, C. E., P. J. Unmack, L. Matsui, and N. Rifenbark. 2007. Comparative phylogeography of five symatric Hypseleotric species (Teleostei: Eleotridae) in south-eastern Australia reveals a complex pattern of drainage basin exchanges with little congruence across species. Journal of Biogeography 34:1518-1533.
Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882.
Turner, T. F., M. V. McPhee, P. Campbell, and K. O. Winemiller. 2004. Phylogeography and intraspecific genetic variation of prochilodontid fishes endemic to rivers of nothern South America. Journal of Fish Biology 64:186-201.
Verheyen, E., W. Salzburger, J. Snoeks, and A. Meyer. 2003. Origin of the superflock of cichlid fishes from Lake Victoria, East Africa. Science (Washington D C) 300:325-329.
Wahlund, S. 1928. Zusammensetzung von Population und Korrelationserscheinung vom Standpunkt der Vererbungslehre aus betrachtet. Hereditas (Lund) 11:65-106.
Wakeley, J., and J. Hey. 1997. Estimating ancestral population parameters. Genetics 145:847-855.
Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theoretical Population Biology 7:256-276.

Weitzman, S. H., and M. Weitzman. 1982. Biogeography and Evolutionary Diversification in Neotropical Freshwater Fishes, with Comments on the Refuge Theory *in* G. T. Prance, ed. Biological Diversification in the Tropics.

Wenzel, J. W., and M. E. Siddall. 1999. Noise. Cladistics 15:51-64.

Whelan, S., P. Liò, and N. Goldman. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends in Genetics 17:262-272.

Whittall, J. B., and S. A. Hodges. 2007. Pollinator shifts drive increasingly long nectar spirs in columbine flowers. Nature 447:406-409.

Wiens, J. J., and T. A. Penkrot. 2002. Delimiting species using DNA and morphological variation and discordant species limits in spiny lizards (Sceloporus). Syst Biol 51:69-91.

Willis, S. C., M. S. Nunes, C. G. Montana, I. P. Farias, and N. R. Lovejoy. 2007. Systematics, biogeography, and evolution of the neotropical peacock basses Cichla (Perciformes : Cichlidae). Molecular Phylogenetics and Evolution 44:291-307.

Willis, S. C., M. S. Nunes, C. G. Montana, I. P. Farias, G. Orti, and N. R. Lovejoy. 2010. The Casiquiare River acts as a corridor between the Amazonas and Orinoco River basins: biogeographic analysis of the genus Cichla. Molecular Ecology 19:1014-1030.

Winemiller, K. O. 2001. Ecology of peacock cichlids (*Cichla* spp.) in Venezuela. Journal of Aquariculture & Aquatic Sciences 9:93-112.

Winemiller, K. O., and D. B. Jepsen. 1998. Effects of seasonality and fish movement on tropical river food webs. Journal of Fish Biology 53:267-296.

Winemiller, K. O., H. López-Fernández, D. C. Taphorn, L. Nico, and A. Barbarino-Duque. 2008. Fish assemblages of the Casiquiare River, a corridor and zoogeographical filter for dispersal between the Orinoco and Amazon basins. Journal of Biogeography 35:1551-1563.

Winemiller, K. O., and S. C. Willis. 2010. Biogeography of the Vaupes Arch and Casiquiare River: Barriers and Passages between the Amazon and Orinoco. Pp. 225-242 *in* J. Albert, and R. E. Reis, eds. Historical Biogeography of Neotrpical Freshwater Fishes. University of California Press, Berkeley, CA.

Wirtz, P. 1999. Mother species-father species: unidirectional hybridization in animals with female choice. Anim Behav 58:1-12.

Won, Y. J., Y. Wang, A. Sivasundar, J. Raincrow, and J. Hey. 2006. Nuclear gene variation and molecular dating of the cichlid species flock of Lake Malawi. Molecular Biology and Evolution 23:828-837.

Wright, S. 1943. Isolation by distance. Genetics 28:114-138.

Zardoya, R., D. M. Vollmer, C. Craddock, J. T. Streelman, S. A. Karl, and A. Meyer. 1996. Evolutionary conservation of microsatellite flanking regions and their use in resolving the phylogeny of cichlid fishes (Pisces: Perciformes). Proceedings of the Royal Society of London B 263:1589-1598.

Zhang, D. X., and G. M. Hewitt. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol Ecol 12:563-584.

Appendices

Chapter 2

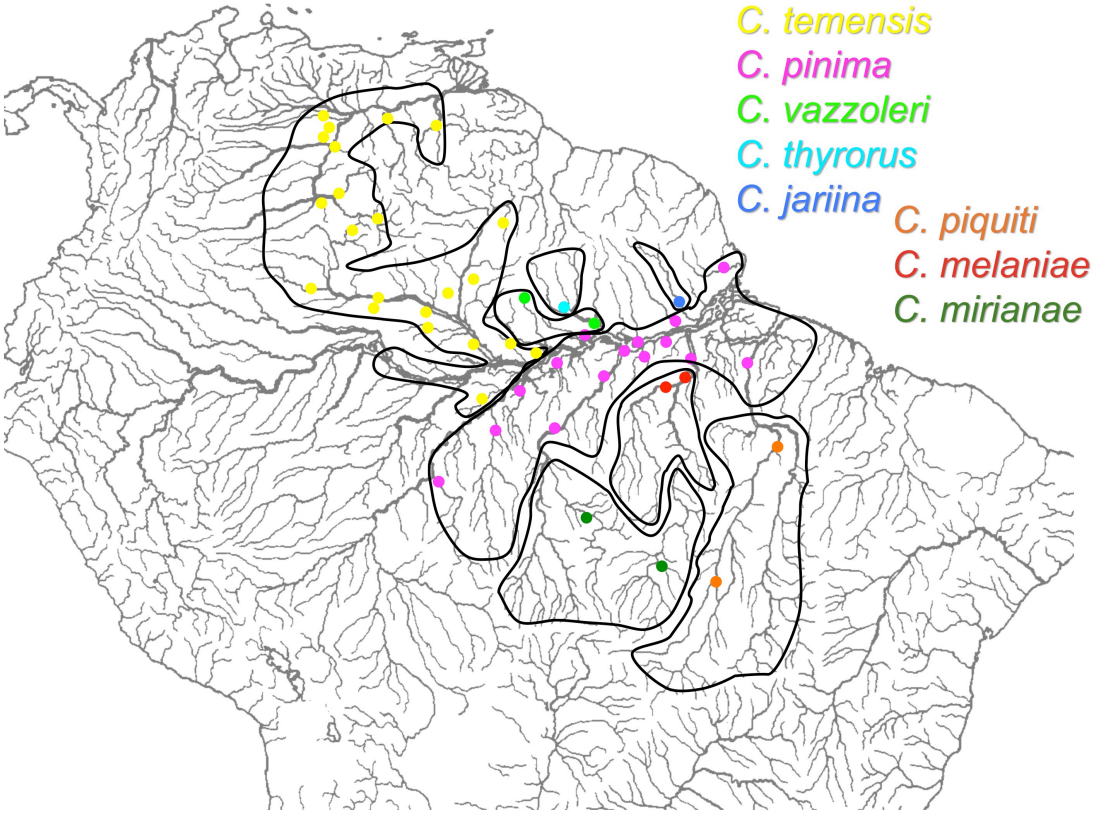
Table S1. Approximate coordinates and Atlantic versant of the localities sampled by the authors. For other sites, see Renno et al. (2006).

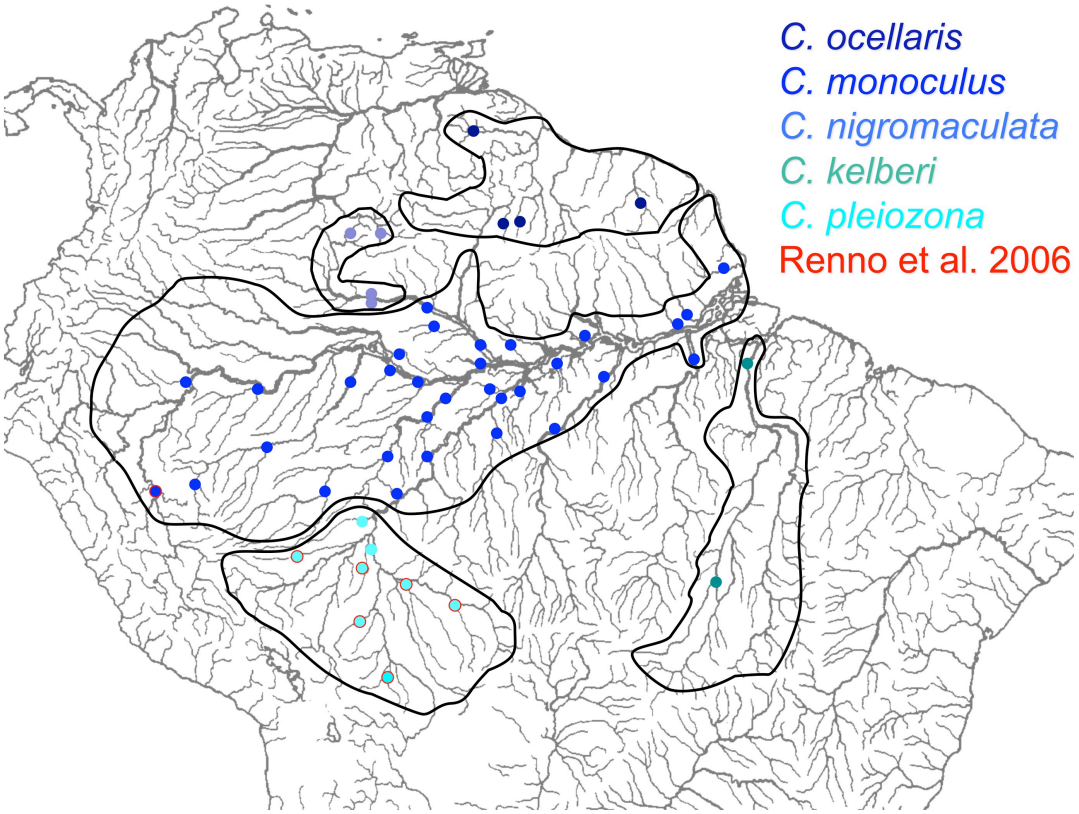
	Locality	Coordinates	River Drainage
TI	Tigre	9° 22.94' N, 63° 0.73' W	Orinoco
GU	Guanipa	9° 34.29' N, 63° 10.32' W	Orinoco
BJ	Buja	9° 33.12' N, 62° 42.71' W	Orinoco
GR	Guri Reservoir (Caroní)	6° 49.69' N, 63° 19.82' W	Orinoco
SI	Sipao	7° 34' N, 65° 13' W	Orinoco
CA	Caura	6° 48.43' N, 64° 49.67' W	Orinoco
CV	Cunavichito	7° 15.04' N, 67° 44.72' W	Orinoco
CP	Capanaparo	6° 53.78' N, 67° 19.5' W	Orinoco
CI	Cinaruco	6° 32.36' N, 67° 21.84' W	Orinoco
PZ	Parguaza	6° 24.39' N, 67° 10.05' W	Orinoco
AT	Atabapo	3° 46.86' N, 67° 37.73' W	Orinoco
VE	Ventuari	4° 8.03' N, 66° 37.51' W	Orinoco
OR	Orinoco	3° 18.74' N, 66° 36.58' W	Orinoco
IG	Iguapo	2° 57.7' N, 65° 14.14' W	Orinoco
OC	Ocamo	2° 45.33' N, 65° 1.87' W	Orinoco
MV	Mavaca	2° 26.27' N, 65° 8.72' W	Orinoco
CR	Curamoni	2° 36.67' N, 66° 9.44' W	Amazonas
PA	Perro de Agua	2° 48.7' N, 66° 3.95' W	Amazonas
CQ	Casiquiare	2° 38.21' N, 66° 13.39' W	Amazonas
PS	Pasiba	2° 24.15' N, 66° 26.15' W	Amazonas
UA	Uaupes	0° 12' N, 68° 1.16' W	Amazonas
IM	Ia-mirim	0° 23.37' S, 66° 38.89' W	Amazonas
TE	Teá	0° 30.45' S, 65° 8.26' W	Amazonas
MR	Marauiá	0° 23.1' S, 65° 12.43' W	Amazonas
UE	Uneiuxi	0° 50.15' S, 65° 34.71' W	Amazonas
DA	Daraá	0° 27.07' S, 64° 45.62' W	Amazonas
PT	Preto	0° 14.77' S, 64° 5.93' W	Amazonas
BC	Barcelos	0° 57.99' S, 62° 55.75' W	Amazonas
PI	Pirara (Takutu)	3° 39.4' N, 59° 31.69' W	Amazonas
ES	Essequibo (Rupununi)	3° 42.97' N, 59° 19.1' W	Essequibo
CY	Cuyuni	6° 44.17' N, 61° 7.84' W	Essequibo
MA	Maroni	3° 40.11' N, 54° 5.87' W	Maroni
XE	Xeruini	0° 37.06' S, 61° 57.78' W	Amazonas
TA	Tapera	0° 26.94' N, 61° 27.05' W	Amazonas
UN	Unini	1° 39.08' S, 63° 41.56' W	Amazonas
NA	Novo Airão	2° 37.18' S, 60° 57.15' W	Amazonas
PE	Preta da Eva	2° 56.43' S, 59° 31.2' W	Amazonas
UR	Urubu	3° 7.08' S, 58° 34.67' W	Amazonas
IQ	Iquitos	3° 41.5' S, 73° 15.84' W	Amazonas
TB	Tabatinga	4° 15.15' S, 69° 56.29' W	Amazonas
JA	Juruá (Carauari)	4° 52.97' S, 66° 53.75' W	Amazonas
EI	Eirunepé	6° 39.6' S, 69° 52.43' W	Amazonas
CS	Cruzeiro do Sul	7° 37.85' S, 72° 40.2' W	Amazonas
AM	Lago Amaná	2° 41.35' S, 64° 37.17' W	Amazonas
TF	Tefé	3° 21.24' S, 64° 42.65' W	Amazonas
CO	Coari	4° 5.1' S, 63° 8.45' W	Amazonas
PP	Piagaçu-Purus	4° 26.12' S, 62° 4.85' W	Amazonas
TP	Tapauá	5° 37.67' S, 63° 10.97' W	Amazonas
LB	Labrea	7° 15.53' S, 64° 47.87' W	Amazonas
BA	Boca do Acre	8° 45.12' S, 67° 23.87' W	Amazonas
MC	Manacapuru	3° 17.98' S, 60° 37.23' W	Amazonas
IA	Igapo-Açu	4° 32.74' S, 60° 48.51' W	Amazonas
BO	Borba	4° 23.25' S, 59° 35.74' W	Amazonas
CM	Canumã	4° 0.49' S, 59° 6.01' W	Amazonas
AP	Aripuanã	5° 10.24' S, 60° 22.39' W	Amazonas
HU	Humaita	7° 30.37' S, 63° 1.25' W	Amazonas
MD	Machado	8° 4.42' S, 62° 53.5' W	Amazonas
CN	Cunia	8° 19.25' S, 63° 28.08' W	Amazonas
CC	Canaçari	3° 2.67' S, 58° 22.05' W	Amazonas
MS	Maués	3° 23.02' S, 57° 43.12' W	Amazonas
JT	Jatapu (Uatumã)	0° 23.97' N, 59° 26.7' W	Amazonas
NH	Nhamunda	2° 11.15' S, 56° 42.77' W	Amazonas
TS	Terra Santa	2° 6.67' S, 56° 29.39' W	Amazonas
TR	Trombetas (abv. rapids)	1° 3.25' S, 57° 3.72' W	Amazonas
OX	Oriximiná	1° 45.95' S, 55° 51.95' W	Amazonas
LG	Lago Grande	2° 22.12' S, 54° 25.68' W	Amazonas
TL	Tapajós mouth	2° 30.31' S, 54° 57.13' W	Amazonas
IT	Itaituba	4° 16.55' S, 55° 59.04' W	Amazonas
JC	Jacareacanga	6° 13.62' S, 57° 44.67' W	Amazonas
CU	Curuá-Una	2° 51.52' S, 54° 20.92' W	Amazonas

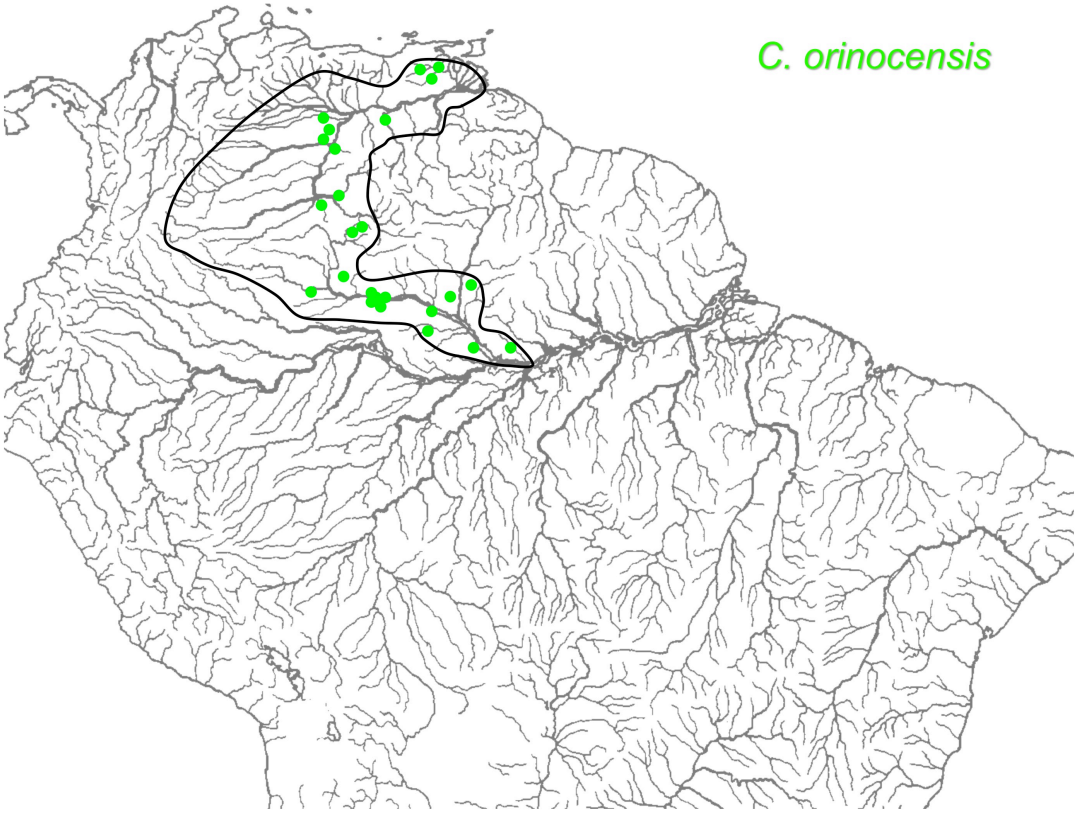
Table S1. *continued*

PU	Paru	1° 14.18' S, 53° 3.26' W	Amazonas
GA	Guajara	1° 49.33' S, 53° 2.27' W	Amazonas
VX	Vittoria do Xingu	2° 52.85' S, 52° 0.95' W	Amazonas
JR	Jari (lower)	0° 49.42' S, 52° 27.51' W	Amazonas
JU	Jari (above waterfalls)	0° 32.61' S, 52° 40.82' W	Amazonas
AR	Araguari	0° 56.01' N, 51° 0.22' W	Amazonas
AF	Alta Floresta	9° 54.32' S, 56° 6.74' W	Amazonas
SM	Suia Missu	11° 44.5' S, 52° 21.36' W	Amazonas
XA	Xingu (Altamira)	3° 15.7' S, 52° 4.48' W	Amazonas
IR	Iriri	4° 22.69' S, 53° 38.89' W	Amazonas
TO	Tocantins (Baião)	3° 8.07' S, 49° 40.56' W	Amazonas
AG	Araguatins	3° 39.05' S, 48° 7.43' W	Amazonas
SF	São Felix do Araguaia	11° 46.56' S, 50° 43.06' W	Amazonas
AB	Abunã	9° 49.12' S, 65° 38.92' W	Amazonas
GM	Guajará-Mirim	10° 47.49' S, 65° 20.95' W	Amazonas

Figure S1. Maps of approximate distributions of the 15 described species of *Cichla*.






C. orinocensis

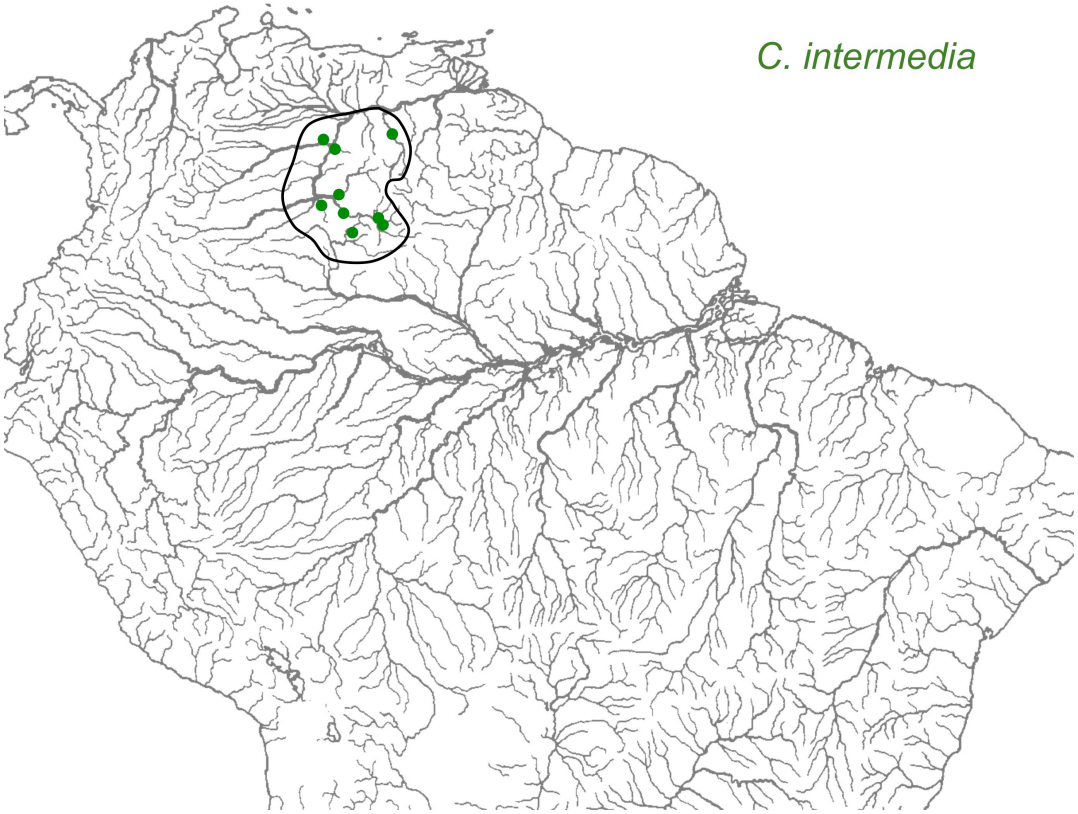
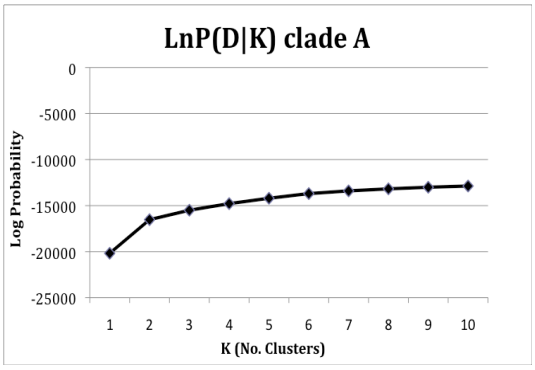
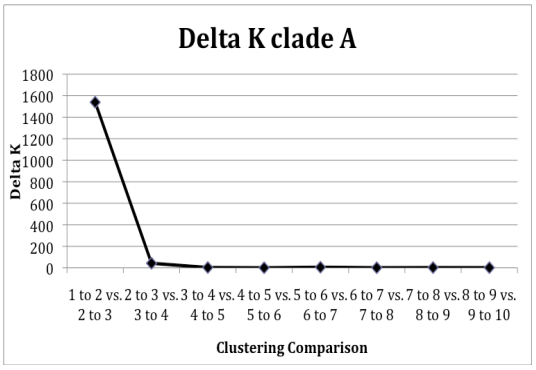
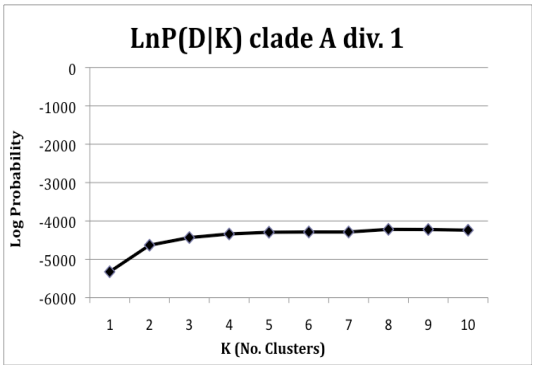
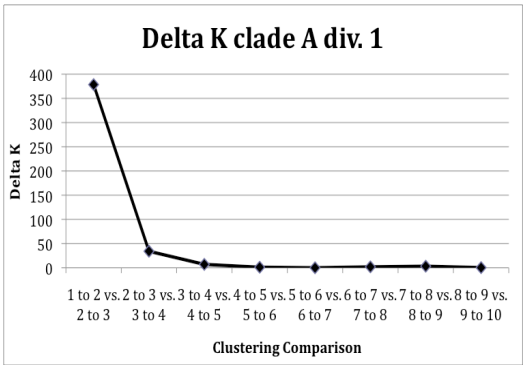

C. intermedia

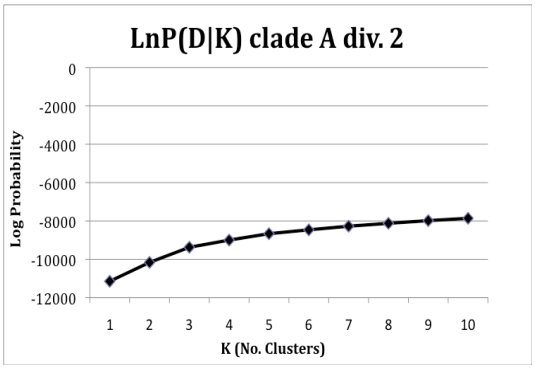
Figure S2. Plots of LnP(D|K) and ΔK for each division in the Divide-and-Reanalyze STRUCUTRE analyses.

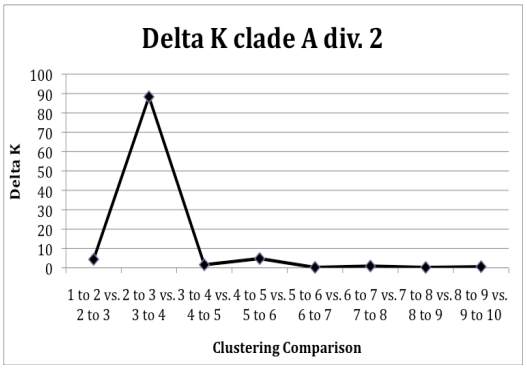


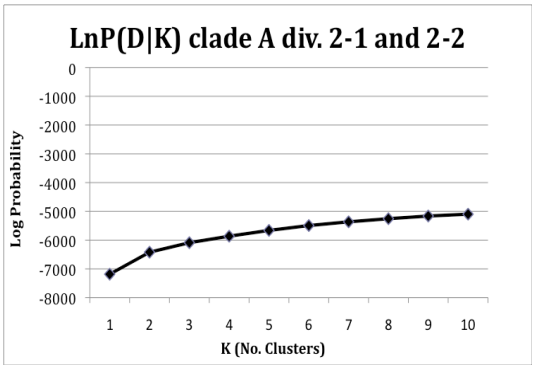


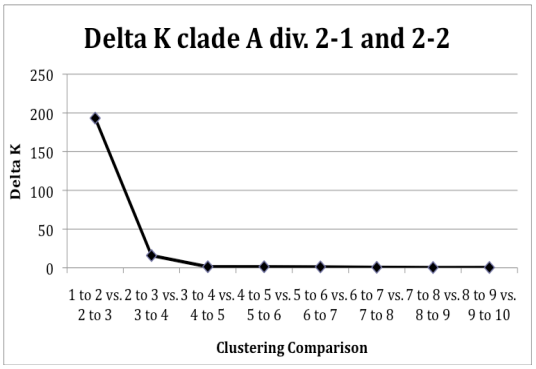


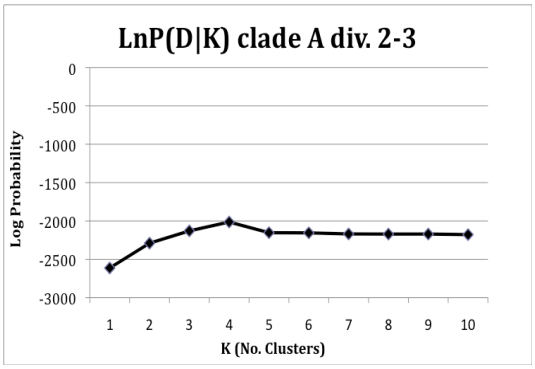


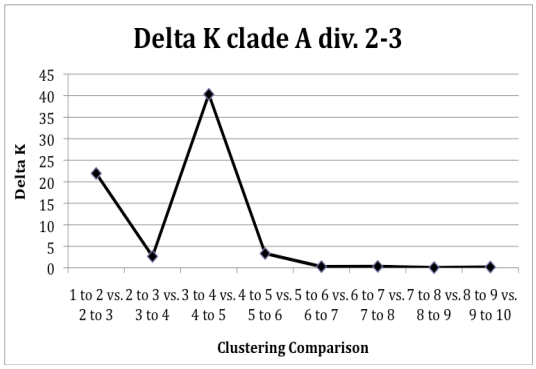


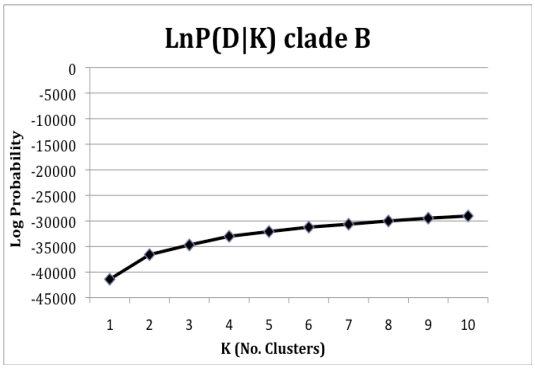


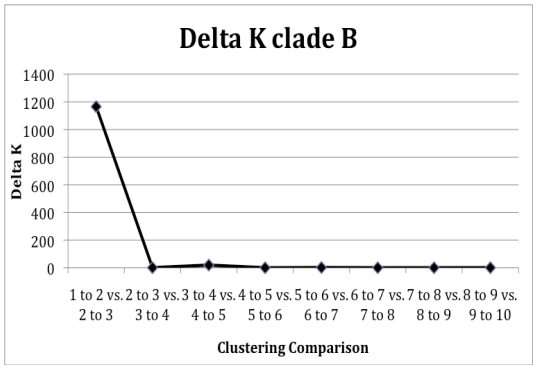


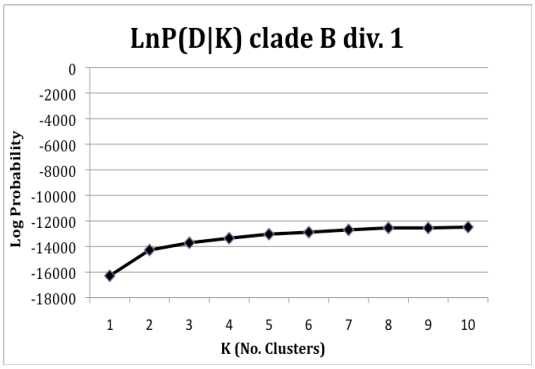


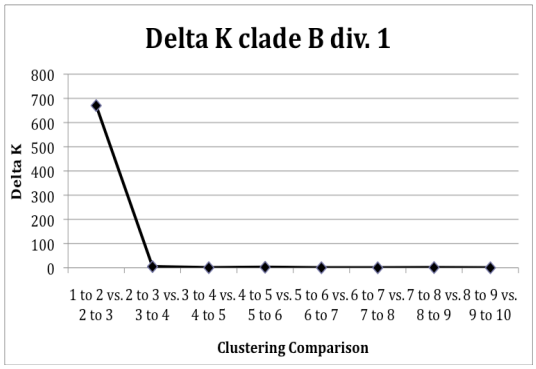

LnP(D|K) clade A div. 2-3
Log Probability
0
-500
-1000
-1500
-2000
-2500
-3000
1 2 3 4 5 6 7 8 9 10
K (No. Clusters)

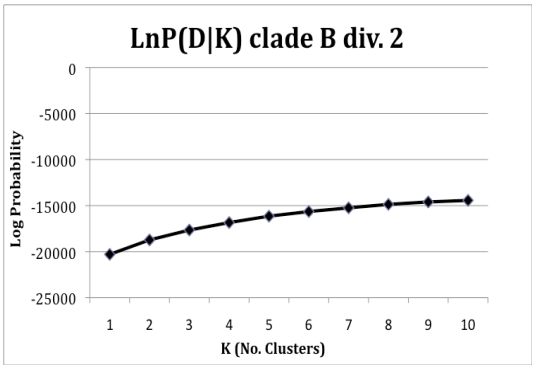

Delta K clade A div. 2-3
Delta K
45
40
35
30
25
20
15
10
5
0
1 to 2 vs. 2 to 3
2 to 3 vs. 3 to 4
3 to 4 vs. 4 to 5
4 to 5 vs. 5 to 6
5 to 6 vs. 6 to 7
6 to 7 vs. 7 to 8
7 to 8 vs. 8 to 9
8 to 9 vs. 9 to 10
Clustering Comparison

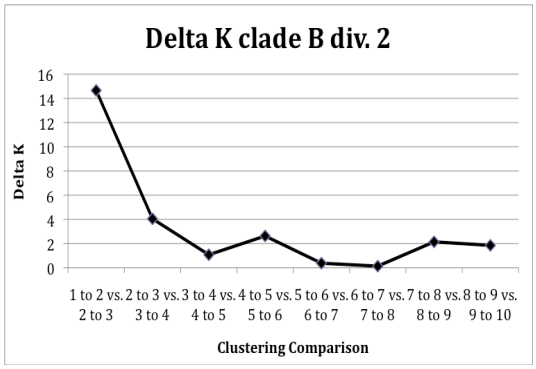

LnP(D|K) clade B
Log Probability
0
-5000
-10000
-15000
-20000
-25000
-30000
-35000
-40000
-45000
1 2 3 4 5 6 7 8 9 10
K (No. Clusters)

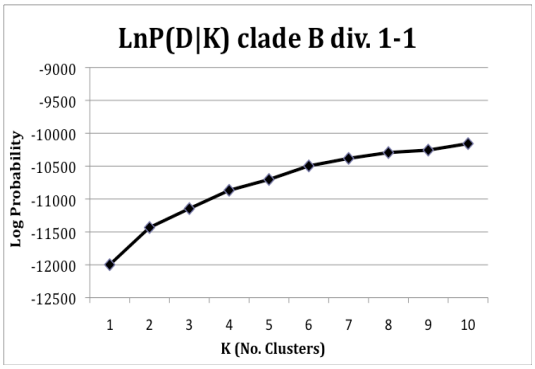

Delta K clade B
Delta K
1400
1200
1000
800
600
400
200
0
1 to 2 vs. 2 to 3
2 to 3 vs. 3 to 4
3 to 4 vs. 4 to 5
4 to 5 vs. 5 to 6
5 to 6 vs. 6 to 7
6 to 7 vs. 7 to 8
7 to 8 vs. 8 to 9
8 to 9 vs. 9 to 10
Clustering Comparison

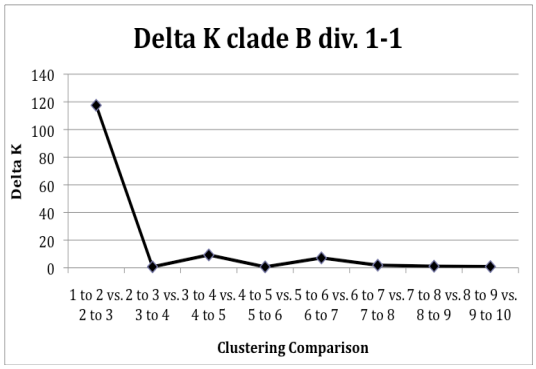

LnP(D|K) clade B div. 1
Log Probability
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
-16000
-18000
1 2 3 4 5 6 7 8 9 10
K (No. Clusters)

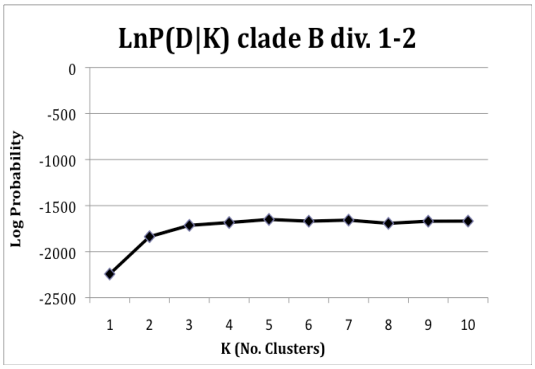

Delta K clade B div. 1
Delta K
800
700
600
500
400
300
200
100
0
1 to 2 vs. 2 to 3
2 to 3 vs. 3 to 4
3 to 4 vs. 4 to 5
4 to 5 vs. 5 to 6
5 to 6 vs. 6 to 7
6 to 7 vs. 7 to 8
7 to 8 vs. 8 to 9
8 to 9 vs. 9 to 10
Clustering Comparison

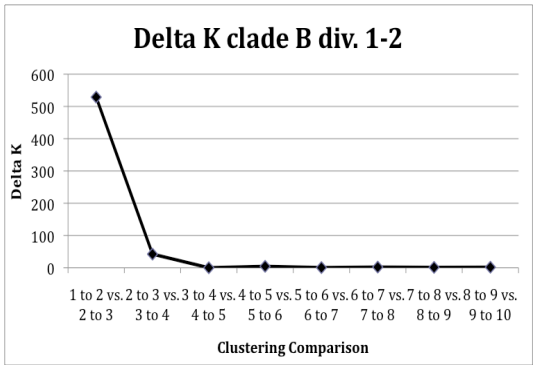

LnP(D|K) clade B div. 2
Log Probability
0
-5000
-10000
-15000
-20000
-25000
1 2 3 4 5 6 7 8 9 10
K (No. Clusters)


Delta K clade B div. 2
Delta K
16
14
12
10
8
6
4
2
0
1 to 2 vs. 2 to 3
2 to 3 vs. 3 to 4
3 to 4 vs. 4 to 5
4 to 5 vs. 5 to 6
5 to 6 vs. 6 to 7
6 to 7 vs. 7 to 8
7 to 8 vs. 8 to 9
8 to 9 vs. 9 to 10
Clustering Comparison









Chapter 3

Supplemental Table 1. PCR primers for loci described by Li et al. (2010) or developed in this study.

marker	reference	forward primer (5'-3')	reverse primer (5'-3')
1835e6	Li et al. 2009	CAGACAAACATCGYTACTGGGARAC	CTTGTCATCCACYATGTTCCAGATGAT
8680e2	"	GATATGGTGGAYTACCTGAACGASCTG	TCCTCAGCKATGTGGTGRATGAA
8680e3	"	GGAGGAGARTTYAAGAAGTAYCTGGACAT	CSCCCTTCAGGCCCTGGATGAT
14867e1	"	CCACAARTACAAGGCCAAGAGRAACTG	GTTCTCCTTSTCCTGSACGGTCTT
18049e2	"	GTGGTGGAGATGCAYGAYGTGAC	TAGTAAAGGTCYCCRTGGATGGTGAG
35564e5	"	AAGACTCAAGRGTGTAYGAGCTGACCAA	CATGTCATCACRTATTCRCTCTTCTGRTT
36298e1	"	GATCCTGAGGGAYTCCCAYGGTGT	GGGCCAGGACTCTCYTGGTCTTGTAGT
55305e1	"	CCTAGTGGACTGTARTAACGCCCCYCT	AAGCCATCCAGTTTGCATAAACACTATC
55378e1	"	ATGARGAAAATGAGGCCAACTTGCT	GCCACCTGKGTATTGATTATAGCTGAG
CorE7	this study	GATGAAGGTTTGGACTTTCCTCT	GGCTACATGACATACAGAGAATGC
CorF12	"	CACACTAAACCAACTTTCTGTGTCA	GCATGTGTCTATTTTCATGTTCTGC
Cpiorcons	"	TTCAGTCTCATCTAATCCCACAAA	CTCCCAACATGGGCACTC
CinKA7	"	GCTCCATCAGTCTCCCAATC	GCCTGGCCTCTTTCTTTCTC
CmoME5	"	TGAAGAAGCAGTTAGCATCACTTT	ACTGCTAAGGTCGTGAAGCTG
CmoMJ3	"	TGACAGTGTTTTTCCACAGCA	TGAACAGTGTGCCTCACAGA
CteOA7	"	CCTGCAGGGTTAGTAGGTTT	GGATGGAGAAGTCCATGTCG
CteOI2	"	TGGATGCAGTATGCAAACACT	AGAGGTGACCACGCTTTGAG

Chapter 4

SI Table 1. Sampling localities and sample sizes for the focal species (control region).

	Locality	Coordinates	Drainage	*C. temensis*	*C. monoculus*	*C. orinocensis*
DL	Orinoco delta	9° 34.29' N, 63° 10.32' W	Orinoco			10
CR	Caroní	6° 49.69' N, 63° 19.82' W	Orinoco	11		
CA	Caura (lower)	7° 34' N, 65° 13' W	Orinoco	10		10
AP	Apure	7° 39.07' N, 66° 31.23' W	Orinoco			8
CN	Cunavichito	7° 15.04' N, 67° 44.72' W	Orinoco	1		3
CP	Capanaparo	6° 53.78' N, 67° 19.5' W	Orinoco	10		10
CI	Cinaruco	6° 32.36' N, 67° 21.84' W	Orinoco	12		9
PR	Parguaza	6° 24.39' N, 67° 10.05' W	Orinoco	2		11
OR	Orinoco (P. Ayacucho)	5° 40' N, 67° 38' W	Orinoco		7	
AT	Atabapo	3° 46.86' N, 67° 37.73' W	Orinoco	10		10
VE	Ventuari	4° 8.03' N, 66° 37.51' W	Orinoco	9		11
IG	Iguapo	2° 57.7' N, 65° 14.14' W	Orinoco	1		
CS	Casiquiare	2° 25.85' N, 66° 24.96' W	Amazonas	10	10	10
UE	Uneiuxi	0° 50.15' S, 65° 34.71' W	Amazonas			12
SI	Santa Isabel	0° 24.84' S, 65° 1.14' W	Amazonas	2	1	1
TP	Tapera	0° 26.94' N, 61° 27.05' W	Amazonas	12		
BA	Barcelos	0° 57.99' S, 62° 55.75' W	Amazonas	5	9	
XE	Xeruini	0° 37.06' S, 61° 57.78' W	Amazonas	7		
UN	Unini	1° 39.08' S, 63° 41.56' W	Amazonas	15	5	
PI	Pirara (Takutu)	3° 39.4' N, 59° 31.69' W	Amazonas	1		
IQ	Iquitos	3° 41.5' S, 73° 15.84' W	Amazonas		4	
TB	Tabatinga	4° 15.15' S, 69° 56.29' W	Amazonas		12	
JU	Juruá	4° 52.97' S, 66° 53.75' W	Amazonas		8	
TF	Tefé + Lago Amaná	3° 21.24' S, 64° 42.65' W	Amazonas		14	
CO	Coari	4° 5.1' S, 63° 8.45' W	Amazonas		4	
PU	Piagacu-Purus	4° 26.12' S, 62° 4.85' W	Amazonas		2	
IA	Igapo-Açu	4° 32.74' S, 60° 48.51' W	Amazonas	10	10	
BO	Borba	4° 23.25' S, 59° 35.74' W	Amazonas		9	
TS	Terra Santa	2° 6.67' S, 56° 29.39' W	Amazonas		10	
TL	Tapajós (lower)	4° 16.55' S, 55° 59.04' W	Amazonas		14	
JR	Jari (lower)	0° 49.42' S, 52° 27.51' W	Amazonas		8	
XL	Xingu (lower)	2° 52.85' S, 52° 0.95' W	Amazonas		10	
AR	Araguari	0° 56.01' N, 51° 0.22' W	Amazonas		2	
Totals:				128	139	105

SI Table 2. Sampling localities and sample sizes for the phylogenetic analysis (cytochrome *b*, ATPase 8,6, and control region).

	Locality	Coordinates	Drainage	*C. temensis*	*C. pinima*	*C. melaniae*	*C. mirianae*	*C. thyrorus*	*C.* cf. *vazzoleri*	*C. jariina*	*C. piquiti*	*C. ocellaris*	*C. monoculus*	*C. pleiozona*	*C. kelberi*	*C. orinocensis*	*C. intermedia*
DL	Orinoco delta	9° 34.29' N, 63° 10.32' W	Orinoco													6	
CR	Caroní	6° 49.69' N, 63° 19.82' W	Orinoco	2													
CL	Caura (lower)	7° 34' N, 65° 13' W	Orinoco													5	
CM	Caura (middle)	6° 48.43' N, 64° 48.67' W	Orinoco														2
CP	Capanaparo	6° 53.78' N, 67° 19.5' W	Orinoco													2	
CI	Cinaruco	6° 32.36' N, 67° 21.84' W	Orinoco	1												3	1
AY	Ayacucho	5° 40' N, 67° 38' W	Orinoco										2				
AT	Atabapo	3° 46.86' N, 67° 37.73' W	Orinoco	1												1	
VE	Ventuari	4° 8.03' N, 66° 37.51' W	Orinoco													2	1
IG	Iguapo	2° 57.7' N, 65° 14.14' W	Orinoco	1													
OC	Ocamo	2° 45.33' N, 65° 1.88' W	Orinoco														1
CY	Cuyuni	6° 44.18' N, 61° 7.84' W	Essequibo									1					
RU	Rupununi	3° 47.92' N, 59° 19.92' W	Essequibo									6					
MA	Maroni	3° 38.72' N, 54° 1.52' W	Maroni									2					
CS	Casiquiare	2° 34.98' N, 66° 24.89' W	Amazonas														1
PS	Pasiba	2° 25.85' N, 66° 24.96' W	Amazonas	1									3			5	
UE	Uneiuxi	0° 50.15' S, 65° 34.71' W	Amazonas													7	
TP	Tapera	0° 26.94' N, 61° 27.05' W	Amazonas	2													
BR	Barcelos	0° 57.99' S, 62° 55.75' W	Amazonas										4				
XE	Xeruini	0° 37.06' S, 61° 57.78' W	Amazonas	3													
UN	Unini	1° 39.08' S, 63° 41.56' W	Amazonas	5									4				
PI	Pirara (Takutu)	3° 39.4' N, 59° 31.69' W	Amazonas	1								3					
IQ	Iquitos	3° 41.5' S, 73° 15.84' W	Amazonas										2				
IA	Igapo-Açu	4° 32.74' S, 60° 48.51' W	Amazonas	2									1				
JT	Jatapu	0° 23.97' N, 59° 26.7' W	Amazonas						5								
AP	Aripuanã	5° 10.24' S, 60° 22.39' W	Amazonas		6												
GM	Guajará-Mirim	10° 47.49' S, 65° 20.95' W	Amazonas											4			
TR	Trombetas (upper)	1° 3.25' S, 57° 3.72' W	Amazonas					2									
OR	Oriximiná	1° 45.95' S, 55° 51.95' W	Amazonas		4												
TM	Tapajós (Alter do Chao)	2° 30.31' S, 54° 57.13' W	Amazonas		2												
TL	Tapajós (Itaituba)	4° 16.55' S, 55° 59.04' W	Amazonas		3												
AF	Alta Floresta	9° 54.32' S, 56° 6.74' W	Amazonas				5										
CU	Curuá-Una	2° 51.52' S, 54° 20.92' W	Amazonas		1												
JC	Jari (upper)	0° 32.61' S, 52° 40.82' W	Amazonas							5							
JR	Jari (lower)	0° 49.42' S, 52° 27.51' W	Amazonas										1				
XL	Xingu (lower)	2° 52.85' S, 52° 0.95' W	Amazonas		1												
XM	Xingu (middle)	3° 15.7' S, 52° 4.48' W	Amazonas			2											
IR	Iriri (middle Xingu)	4° 22.69' S, 53° 38.89' W	Amazonas			5											
SM	Suia Missu (upper Xingu)	11° 44.5' S, 52° 21.36' W	Amazonas				2										
AR	Araguari	0° 56.01' N, 51° 0.22' W	Amazonas		5								1				
SF	São Felix do Araguaia	11° 46.56' S, 50° 43.06' W	Amazonas								5				5		
Totals:				**19**	**22**	**7**	**7**	**2**	**5**	**5**	**5**	**12**	**18**	**4**	**5**	**31**	**6**

SI Table 3. Statistics from the marginal distributions of parameters from the final IMa analyses of *C. temensis*, *C. orinocensis*, and *C. monoculus*. Models are with and without gene flow, and the analyzed basin groupings (see text). Numbers represent model output before conversion using mutation rates and generation times.

Cichla temensis

		run1					run2				
		mean	**HiPt**	**min bin**	**95% low**	**95% high**	**mean**	**HiPt**	**min bin**	**95% low**	**95% high**
Ori vs. Cas + Amaz	**qO**	27922	22858	253	9700	55752	27922	22740	337	9649	55803
	qN	61436	53728	759	29774	108215	61461	53745	911	29656	108401
	qa	248703	44281	253	12399	493168	248703	44281	253	12399	493168
	mNtoO	6.22E-06	4.82E-07	3.71E-08	1.85E-07	2.62E-05	6.02E-06	4.67E-07	2.22E-08	1.56E-07	2.47E-05
	mOtoN	7.21E-06	5.00E-06	3.71E-08	1.15E-06	1.82E-05	7.21E-06	5.05E-06	2.22E-08	1.13E-06	1.82E-05
	t	1.79E+06	6.02E+05	2.19E+04	2.51E+05	3.29E+06	1.79E+06	6.02E+05	2.19E+04	2.51E+05	3.29E+06
	qO	24604	18961	202	7760	52969	24604	18961	202	7760	52969
	qN	115081	74140	1265	40739	348431	115105	74140	1265	40739	348431
	qa	104066	80719	759	44787	220901	104066	80719	759	41414	220901
	t	2.67E+04	1.86E+04	1.26E+03	8.35E+03	5.90E+04	2.67E+04	1.86E+04	1.26E+03	8.35E+03	5.90E+04

		run1					run2				
		mean	**HiPt**	**min bin**	**95% low**	**95% high**	**mean**	**HiPt**	**min bin**	**95% low**	**95% high**
Ori + Cas vs. Amaz	**qO**	32778	27244	422	14086	61825	32768	27454	380	14212	61361
	qN	52093	45125	590	23701	94214	51943	45403	531	23794	92637
	qa	246204	52885	253	12399	492662	328367	50270	337	15857	656883
	mNtoO	5.47E-06	3.00E-06	3.71E-08	4.82E-07	1.60E-05	5.44E-06	3.04E-06	1.48E-08	4.89E-07	1.58E-05
	mOtoN	2.77E-06	3.71E-08	3.71E-08	3.71E-08	1.03E-05	2.74E-06	1.48E-08	1.48E-08	7.41E-08	1.03E-05
	t	1.77E+06	2.81E+06	1.52E+04	1.67E+05	3.30E+06	1.77E+06	2.58E+06	1.18E+04	2.04E+05	3.29E+06
	qO	37959	32304	590	17122	69922	37959	32304	590	17122	69922
	qN	54740	44281	422	23195	108215	54740	44281	422	23195	108215
	qa	159270	106022	253	53897	384362	159270	106022	253	53897	384362
	t	5.16E+04	4.37E+04	8.27E+03	2.21E+04	1.01E+05	5.16E+04	4.37E+04	8.27E+03	2.21E+04	1.01E+05

Cichla orinocensis

		run1					run2				
		mean	**HiPt**	**min bin**	**95% low**	**95% high**	**mean**	**HiPt**	**min bin**	**95% low**	**95% high**
Ori vs. Cas + Amaz	**qO**	79333	71311	1377	42126	131883	79696	71861	1377	42126	132434
	qN	61936	50385	275	25606	122522	62137	50385	275	25606	122522
	qa	357695	81865	367	17988	714758	446275	79387	459	21568	893447
	mNtoO	4.97E-06	3.58E-06	3.41E-08	1.33E-06	1.17E-05	4.96E-06	3.58E-06	3.41E-08	1.40E-06	1.17E-05
	mOtoN	2.27E-06	3.41E-08	3.41E-08	3.41E-08	1.05E-05	2.22E-06	3.41E-08	3.41E-08	3.41E-08	1.04E-05
	t	2.00E+06	2.41E+06	2.75E+04	3.14E+05	3.59E+06	1.97E+06	1.07E+06	2.75E+04	3.18E+05	3.58E+06
	qO	104111	92236	1927	55341	176487	104343	92786	1927	55341	176487
	qN	82418	59747	275	29460	189152	82415	59747	275	29460	189152
	qa	154704	114262	275	56993	351046	154689	114262	275	56443	351597
	t	7.10E+04	4.96E+04	5.51E+03	2.46E+04	1.45E+05	7.17E+04	6.35E+04	3.30E+03	2.61E+04	1.45E+05

		run1					run2				
		mean	**HiPt**	**min bin**	**95% low**	**95% high**	**mean**	**HiPt**	**min bin**	**95% low**	**95% high**
Ori + Cas vs. Amaz	**qO**	85161	77919	1927	44328	140143	85218	78469	1927	44328	139593
	qN	82210	53139	275	22852	219438	82354	53139	275	22852	219438
	qa	356364	94347	367	17988	714758	444328	88565	459	22485	893447
	mNtoO	3.35E-06	3.41E-08	3.41E-08	1.02E-07	1.18E-05	3.29E-06	3.41E-08	3.41E-08	1.02E-07	1.15E-05
	mOtoN	1.05E-05	3.41E-08	3.41E-08	1.70E-07	3.47E-05	1.04E-05	3.41E-08	3.41E-08	1.70E-07	3.39E-05
	t	1.99E+06	1.43E+06	1.65E+04	2.33E+05	3.59E+06	1.98E+06	1.21E+06	1.65E+04	2.62E+05	3.58E+06
	qO	101697	91134	2478	54240	170430	101704	91134	2478	54240	170980
	qN	141809	80121	275	38271	413822	141433	80121	275	38271	412720
	qa	156683	118117	275	58095	346641	156872	118117	275	58645	346641
	t	6.98E+04	4.99E+04	7.38E+03	2.54E+04	1.45E+05	6.96E+04	5.28E+04	4.52E+03	2.52E+04	1.45E+05

Cichla monoculus

		run1					run2				
		mean	**HiPt**	**min bin**	**95% low**	**95% high**	**mean**	**HiPt**	**min bin**	**95% low**	**95% high**
Ori vs. Cas + Amaz	**qO**	15911	55	18	128	35297	13332	18	18	165	35040
	qN	179781	61858	184	24046	357012	175507	58921	184	27349	356645
	qa	709686	131241	918	52313	1759361	689625	134912	918	54148	1755690
	mNtoO	1.06E-03	9.56E-05	2.22E-06	3.78E-05	3.85E-03	1.61E-03	1.11E-05	2.22E-06	3.33E-05	4.35E-03
	mOtoN	1.31E-03	6.67E-06	2.22E-06	1.11E-05	4.20E-03	1.01E-03	1.56E-05	2.22E-06	1.11E-05	3.95E-03
	t	7.72E+05	4.39E+04	3.37E+03	1.69E+04	3.07E+06	6.92E+05	7.25E+04	1.69E+03	2.19E+04	2.99E+06
	qO	7149	55	18	128	29277	7166	92	18	128	29314
	qN	193000	44420	367	23128	687592	192570	43686	367	22394	687592
	qa	144829	62775	367	36344	529736	145235	62775	367	36344	530470
	t	3.46E+04	8.81E+03	1.01E+02	2.73E+03	1.09E+05	3.46E+04	7.59E+03	1.01E+02	2.53E+03	1.09E+05

Supplemental Table 1. Output summary from the BEST analyses of species trees.

mtDNA	clade B1 *C. orinocensis*	harmonic mean LnL	mean branch lengths (95% HPD)				
			C. intermedia/ C. orinocensis	ancestor of *intermedia+orinocensis*	*C. oc. monoculus*	ancestor of *intermedia+monoculus*	*C. temensis*
no	no	-15949.8	0.00164 (0.00078 - 0.00247)	0.00155 (0.00065 - 0.00269)	0.00319 (0.00261 - 0.00378)	0.00117 (0.00063 - 0.00186)	0.00436 (0.00364 - 0.00504)
no	yes	-16042.912	0.00257 (0.00199 - 0.00311)	0.00077 (0.00038 - 0.00141)	0.00334 (0.00291 - 0.0038)	0.00077 (0.00042 - 0.00125)	0.00411 (0.00353 - 0.00473)
yes	no	-17659.3	0.00222 (0.00122 - 0.00327)	0.00234 (0.00117 - 0.00373)	0.00456 (0.00382 - 0.00534)	0.002 (0.00125 - 0.00283)	0.00656 (0.00569 - 0.00747)
yes	yes	-17759.629	0.0026 (0.00184 - 0.00341)	0.00078 (0.00025 - 0.0016)	0.00338 (0.00272 - 0.00417)	0.00273 (0.00045 - 0.00437)	0.00612 (0.00423 - 0.00745)

Supplemental Table 2. Output summary from the BEST analyses of species trees.

mtDNA	clade B1 *C. orinocensis*	mean population size (Θ) (95% HPD)					
		C. intermedia	*C. orinocensis*	ancestor of *intermedia+orinocensis*	*C. oc. monoculus*	ancestor of *intermedia+monoculus*	MRCA of *Cichla*
no	no	0.00353 (0.00153 - 0.00679)	0.00488 (0.0022 - 0.00924)	0.00142 (0.00047 - 0.00353)	0.00938 (0.0056 - 0.01532)	0.00102 (0.00048 - 0.00191)	0.00056 (0.00029 - 0.00098)
no	yes	0.00556 (0.00239 - 0.00985)	0.01339 (0.00744 - 0.0241)	0.0007 (0.00034 - 0.00143)	0.00818 (0.00467 - 0.01317)	0.00067 (0.00035 - 0.00119)	0.00056 (0.0003 - 0.00092)
yes	no	0.0041 (0.00195 - 0.00793)	0.00537 (0.00268 - 0.00966)	0.00176 (0.00052 - 0.0043)	0.01192 (0.00745 - 0.01867)	0.00159 (0.00078 - 0.00282)	0.00057 (0.00028 - 0.0012)
yes	yes	0.00424 (0.00196 - 0.0085)	0.01258 (0.00742 - 0.02113)	0.00126 (0.00049 - 0.00274)	0.01124 (0.00662 - 0.01836)	0.00361 (0.00065 - 0.00659)	0.00147 (0.00028 - 0.00591)