Comparative Genomics Reveals Adaptive Protein Evolution and a Possible Cytonuclear Incompatibility between European and American Eels

Pierre-Alexandre Gagnaire,*1 Eric Normandeau,1 and Louis Bernatchez1

1Département de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Québec, Quebec, Canada G1V 0A6
*Corresponding author: E-mail: pierre-alexandre.gagnaire.1@ulaval.ca.
Associate editor: Michael Nachman

Abstract

During the early stages of speciation, interspecific gene flow may be impeded by deleterious epistatic interactions in hybrids, which maintain parental allelic combinations at the speciation genes. The resulting semipermeable nature of the barrier to interspecific gene flow provides a valuable framework to identify the genes involved in hybrid mortality or sterility, as well as the evolutionary mechanisms that initially caused their divergence. The two Atlantic eels Anguilla anguilla and A. rostrata are partially isolated sister species that naturally hybridize, but whose genetic basis of postzygotic isolation remains unknown. We collected high-throughput sequencing data from the transcriptomes of 58 individuals and discovered 94 genes showing differentially fixed mutations between species. Evidence for positive selection at nuclear diagnostic genes was obtained using multilocus extensions of the McDonald–Kreitman test with polymorphism data from each species. In contrast, mitochondrial protein-coding genes experienced strong purifying selection and mostly diverged at synonymous sites, except for the mt-atp6 gene, which showed an atypically high nonsynonymous to synonymous rate ratio. Nuclear-encoded protein interactors of the mt-atp6 gene in the ATP synthase complex were significantly overrepresented in the list of nuclear diagnostic genes. Further analysis of resequencing data showed that positive selection has operated at both the mt-atp6 gene and its nuclear interactor atp5c1. These findings suggest that a cytonuclear incompatibility caused by a disruption of normal ATP synthase function in hybrids contributes to partial reproductive isolation between European and American eels.

Key words: Atlantic eels, ATP synthase, genome scan, next generation sequencing, positive selection, speciation genes.

Introduction

A central goal of speciation genetics is to assess the relative importance of adaptive and nonadaptive processes involved in species divergence (Orr and Presgraves 2000; Coyne and Orr 2004; Wu and Ting 2004; Presgraves 2010). This is particularly relevant to understanding the early stages of speciation, when only small numbers of genes are responsible for hybrid dysfunction. The molecular mechanisms underlying genetic incompatibilities have been well described for some speciation genes found in model organisms (e.g., Presgraves et al. 2003; Brideau et al. 2006; Lee et al. 2008; Mihola et al. 2009; Phadnis and Orr 2009; Chou et al. 2010). As predicted by the Dobzhansky–Muller model of hybrid incompatibilities (Dobzhansky 1936; Muller 1942; Orr 1995), these genes often appear to be involved in deleterious epistatic interactions caused by the combination of incompatible alleles from different loci in hybrids. Studies investigating speciation genes have shown that the evolution of incompatible alleles causing postzygotic isolation may be driven either by nonadaptive or by adaptive processes. Moreover, in many cases where positive selection has been evidenced, adaptive evolution has been frequently attributed to compensatory changes initially triggered by mutation or molecular arms races and only rarely to ecological adaptation (Presgraves 2010; Maheshwari and Barbash 2011). Although considerable progress has been made toward characterizing the evolutionary forces underlying the establishment of incompatible alleles, most speciation genes were identified from a small number of model organisms. Therefore, extending speciation genetic approaches to nonmodel species may provide a more comprehensive overview of the mechanisms at play in nature, in particular with respect to their dependence on environment.

Owing to the recent progress in next generation sequencing (NGS) technologies, nonmodel organisms become increasingly amenable to genome-wide investigations (Gilad et al. 2009; Wolf et al. 2010; Ekblom and Galindo 2011). These techniques provide appealing new opportunities to discover genetic incompatibilities in taxa that are still in incipient stages of reproductive isolation and continue to exchange genes via introgressive hybridization (Barreto et al. 2011). Partially reproductively isolated species present a major potential for speciation geneticists: if the level of introgression is sufficient to prevent genetic divergence at genes that do not impair the survival or fertility of hybrids, it may in turn reveal genetic barriers that maintain elevated values of interspecies differentiation or even differentially fixed alleles.

Genomic regions under divergent selection are traditionally identified using FST scan methods, which ideally require
complementary analyses of the molecular signature left by selection on linked neutral variation (Akley 2002; Turner et al. 2008, 2010; Kolaczkowski et al. 2011) or of differential introgression in hybrid zones (Payseur 2010). Another appealing approach is to estimate the proportion of nonsynonymous substitutions fixed by positive selection, by contrasting divergence among species with polymorphism within species (Eyre-Walker 2006). Several multilocus approaches extending the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991) have been developed for this purpose (Fay et al. 2001; Smith and Eyre-Walker 2002; Biern and Eyre-Walker 2004; Eyre-Walker and Keightley 2009). However, these tests have rarely been applied to recently diverged species, and additional investigations are necessary to further test the reliability of MK-related approaches to investigating the early stages of speciation (Gossmann et al. 2010).

The two Atlantic eel species, Anguilla rostrata and A. anguilla, provide a fascinating experimental model for studying the genetic basis of reproductive isolation in the wild. Their spawning areas partially overlap in the Sargasso Sea, from where after hatching, planktonic leptocephali larvae drift with the currents to the North-eastern American coasts for A. rostrata and the European and North African coasts for A. anguilla. These sister species can be distinguished both morphologically and genetically, but the finding of hybrids in Iceland has proved reproductive isolation to be still incomplete (Avise et al. 1990). Although a direct measure of hybrid fitness is technically challenging for eels, first generation hybrids were shown to be viable and fertile based on the genetic identification of backcross genotypes in nature (Albert et al. 2006). However, departures from neutral introgression found at some highly divergent loci suggest that particular allelic combinations may affect hybrid survival at a few genetic incompatibility loci (Gagnaire et al. 2009). While supporting the existence of a semipermeable genetic barrier to interspecific gene flow (Harrison 1986), these findings based on anonymous markers did not identify genes involved in hybrid breakdown. On the other hand, they allowed estimating a very low background level of neutral differentiation between species (average neutral $F_{ST} = 0.07$) and a small proportion of markers with abnormally high levels of divergence. This heterogeneous genomic divergence pattern provides a valuable framework to identify the genes involved in hybrid dysfunction (Nosil et al. 2009) and study the evolutionary mechanisms that initially caused their divergence. Using a high-throughput genome scan approach based on RNA-Seq data, we characterized patterns of polymorphism and divergence across coding genes of the two Atlantic eel species. We show that population genetic and molecular evolution analyses can be efficiently combined to identify genes possibly involved in the early stages of postzygotic reproductive isolation.

Materials and Methods

CDNA Sequence Data
Transcriptome sequencing data (RNA-Seq) generated by the 454 GS-FLX sequencing platform were retrieved from two NCBI sequence read archives. The data set of the American eel A. rostrata (SRA045712) consisted of 940,263 individually tagged reads obtained from 40 specimens evenly sampled from the St Lawrence estuary and Florida (Gagnaire et al. 2012). The data set of the European eel A. anguilla (SRA020995) was obtained by sequencing a pooled cDNA library constructed from 18 individuals evenly sampled from the Vilaine, Tiber, and Sele estuaries (Coppe et al. 2010). The sequencing coverage of A. anguilla was improved by resequencing its library on an additional half plate of 454 GS-FLX. As a result, 371,172 new reads were added to the 310,079 archived reads, representing a total of 681,251 reads for A. anguilla. In both species, whole glass eels were used for RNA extraction.

Sequence Assembly, Data Preparation, and Single Nucleotide Polymorphism Discovery

CLC Genomic Workbench 3.7 (CLC bio) software was used to perform a de novo assembly of the whole sequence data. After several preliminary runs, assembly parameters were set to a minimal read length fraction of 0.5 and a similarity threshold of 0.98, which offered a good compromise between the level of polymorphism, the sequencing error rate, and the need to minimize the inclusion of nonorthologous sequences within contigs.

We used the Neighborhood Quality Standard (NQS) algorithm (Altshuler et al. 2000; Brockman et al. 2008) to discover single nucleotide polymorphisms (SNPs) from aligned reads within contigs, while taking into account the quality values of adjacent bases in order to distinguish sequencing errors from actual SNPs. We set a rare variant frequency threshold of 5% (or a count threshold of 5 for the rarest variant when the coverage exceeded 100×) as well as a minimum coverage of 20× per polymorphic site to avoid the artificial detection of sequencing errors as SNPs. Only biallelic SNPs were considered.

Detection of Diagnostic Mutations
For the A. rostrata data set, each sequence read was assigned to an individual on the basis of its 10 bp bar code unique to each individual. Individual genotypes were then inferred at each polymorphic site using the probabilistic framework described in Gagnaire et al. (2012) and were used to calculate allele frequencies. Because sequence read data for A. anguilla were obtained by sequencing non–bar-coded individual libraries pooled in equal amounts (Coppe et al. 2010), allele frequencies in this species were directly estimated from allele counts at each SNP position. We used a sequencing coverage threshold of 8× per species to remove SNPs covered by less than eight different individuals in A. rostrata and eight different reads in A. anguilla. The allelic frequency differential between species ($\delta_{Aro/Aan}$) was then calculated at each retained SNP, in order to estimate the genome-wide distribution of between-species genetic differentiation and identify diagnostic mutations. Contigs containing SNPs with allele frequency differentials above 0.95 ($\delta_{Aro/Aan} > 0.95$) were all manually inspected in order to detect potential local assembly errors causing the nondetection of diagnostic mutations.
Open Reading Frame Prediction and Sequence Annotation

For each contig containing at least one diagnostic mutation, the consensus sequence was blasted against the non-redundant NCBI protein database (nr) using BLASTX with an E value threshold of $10^{-5}$ (Altschul et al. 1997). Protein-coding regions (open reading frames [ORFs]) within contigs showing significant hits to the nr database were identified with OrfPredictor (Min et al. 2005), using the reading frame information contained in the BLASTX output. For each successfully annotated contig, noncoding regions were trimmed and the coding region was arranged according to its reading frame. Orthologous sequences from at least three other divergent fish species (most frequently Danio rerio, Salmo salar, Tetraodon nigroviridis, Osmerus mordax, Oncorhynchus mykiss, Ictalurus punctatus, and Esx lucius) were retrieved from BLASTX outputs to be used as outgroups. We inferred the most parsimonious ancestral state at each polymorphic site in eels by aligning orthologous genes based on their translated sequences using CLUSTALW (Thompson et al. 1994) in MEGA 5 (Tamura et al. 2011). Polymorphic sites were then classified as either synonymous ($) or nonsynonymous (NS) and derived allele counts were determined in each species and assigned to one of four different frequency classes (low: [0;0.05], moderate: [0.05;0.2], common: [0.2;0.5], and high: [0.5;1]). Functional overrepresentation of Gene Ontology categories in the list of nuclear diagnostic genes was tested with reference to the whole list of genes retained for SNP discovery, using Fisher’s exact test in BLAST2GO v.2 (Götz et al. 2008).

Tests of Molecular Evolution

Because A. rostrata and A. anguilla have most likely diverged less than 500,000 years ago (Gagnaire P-A, unpublished data), most diagnostic genes carry few differentially fixed mutations between the two species. This problem added to a potentially low number of polymorphisms segregating within each species at some genes may cause a large variance in single gene estimates of the proportion of amino acid substitutions driven by positive selection (Bierne and Eyre-Walker 2004). We thus inferred the fraction of adaptive substitutions ($\omega$) in each species using three different multilocus extensions of the MK test (McDonald and Kreitman 1991) implemented in the DofE software available on A. Eyre-Walker Lab’s Web site http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker (University of Sussex, United Kingdom). All of these approaches assume that synonymous mutations are neutral (i.e., selection on synonymous codon usage and linkage effects can be neglected), whereas nonsynonymous mutations are either strongly deleterious, neutral, or strongly advantageous. Under these assumptions, positively selected replacement mutations are expected to go rapidly to fixation and contribute only little to polymorphism compared with divergence between species. The first method (Fay et al. 2001) simply sums the values of $D_{\mu} D_{\rho} P_{\omega}$ and $P_{\omega}$ across genes to estimate $\omega$. However, it is sensitive to the existence of slightly deleterious mutations (SDMs) segregating at low frequencies and to variations in $N_e$ across genes, which usually lead to underestimate $\omega$ (Fay et al. 2001; Bierne and Eyre-Walker 2004; Charlesworth and Eyre-Walker 2008). The second method (Smith and Eyre-Walker 2002) estimates the average fraction of adaptive substitutions by averaging statistics across genes but may be sensitive to the presence in the data set of genes showing little or no polymorphism. The third method (Bierne and Eyre-Walker 2004) estimates $\omega$ within a maximum likelihood (ML) framework, which allows including genes that have little or no polymorphism. Confidence intervals of $\omega$ were obtained by bootstrapping genes for the first two methods and by determining the $2 \times$ units of Log($L$) interval for the third one. Because we did not sample the same number of chromosomes for all genes, we did not use the method that estimates the distribution of fitness effects of new mutations to account for the presence of SDMs (Eyre-Walker and Keightley 2009). However, we controlled for the effects of negative selection acting on SDMs by removing rare polymorphisms segregating at low frequencies ($\leq 20\%$) in each species, as commonly applied in other studies (Fay et al. 2001, 2002; Bierne and Eyre-Walker 2004; Charlesworth and Eyre-Walker 2008; Axelson and Ellegren 2009; Gossmann et al. 2010). Since both nuclear and mitochondrial genes were found in the list of diagnostic genes, we estimated $\omega$ separately from nuclear and mitochondrial gene sub-data sets, using alternatively A. rostrata or A. anguilla as the ingroup species to score polymorphism. For each species, two published sequences available for each of the 13 mitochondrial protein-coding genes (Minagishi et al. 2005) were retrieved from GenBank. Short 454 reads mapping to these reference sequences were used to improve the identification of divergent sites between species. The number of nonsynonymous substitutions per nonsynonymous site ($d_{ns}$), the number of synonymous substitutions per synonymous site ($d_{sy}$), and the $\omega$ ratio ($d_{ns}/d_{sy}$) were estimated for each gene using DnaSP v5 (Librado and Rozas 2009).

Resequencing of Candidate Genes

We resequenced two genes showing relatively high ratios of nonsynonymous to synonymous divergence (see Results), which are involved in the ATP biosynthetic process: the mitochondrial gene mt-atp6 and the nuclear gene atp5c1, coding for two interacting subunits of the cytonuclear protein complex ATP synthase (OXPHOS complex V). The mt-atp6 gene encodes the ATP synthase F0 subunit 6, a key component of the proton channel located in the inner mitochondrial membrane, whereas the atp5c1 gene encodes the ATP synthase F1 subunit gamma, which connects the F0 rotary motor to the F1 catalytic core of the ATP synthase. New primers were designed to resequence these genes in 20 American eels (A. rostrata) from the St. Lawrence estuary and 20 European eels (A. anguilla) from the Vilaine estuary. The full length–coding sequence of the mt-atp6 gene (684 bp) was amplified using primers ANG-ATP6-L (5'-GTATTCTCATGGGCGGTGTT-3') and ANG-ATP6-R (5'-CATGGGCTTGGGTCAACTAT-3'). An exon priming–intron crossing approach was developed to amplify overlapping
fragments covering the whole \textit{atp5c1} gene for two individuals from each species. Two overlapping fragments located in the most diverging region of \textit{atp5c1} and covering 1,627 bp from intron 5 to exon 8 were amplified in all 40 individuals using the primer pairs ANG-ATP5C1-5L (5'-TGTTCTGATTCAACCTGATT-3') and ANG-ATP5C1-6R (5'-CATGCAGGTAAGCTGACTATC-3') as well as ANG-ATP5C1-6L (5'-CTGCAGGTAAGCTGACTATC-3') and ANG-ATP5C1-8R (5'-AGAGCAGCTGACCAGAGAT-3'). Polymerase chain reactions (PCRs) were performed in 20 μl reaction volumes containing 1 × PCR buffer, 1.5 mM of MgCl₂, 0.25 mM of each dNTP, 0.5 μM of forward and reverse primers, 0.5 unit of GoTaq DNA polymerase (Promega), and 100 ng of purified genomic DNA. The amplification parameters were as follows: 94 °C for 5 min followed by 35 cycles of (94 °C for 30 s, 60 °C for 30 s, 72 °C for 45 s), and ended by 4 min at 72 °C. Amplicons were purified with ExoSAP-IT (USB Corporation) and sequenced in both directions on an ABI 3130 xi Genetic Analyzer (Applied Biosystems), using the amplification primers and the Big Dye Terminator Sequencing Kit (v3.1 Applied Biosystems). Sequences were manually inspected in BioEdit and aligned using CLUSTALW (Thompson et al. 1994). Heterozygous positions at the \textit{atp5c1} locus, identified by double peaks in individual electropherograms, were represented by their IUPAC ambiguity code. Individual haplotypes were then inferred for individual sequences containing ambiguities (at least two heterozygous SNPs) using the ISHAPE software (Delaneau et al. 2007).

Population genetic statistics, including the number of segregating sites (S), the number of haplotypes (h), the average number of nucleotide differences (k), nucleotide diversity (π) (Nei 1987), and Watterson’s estimator of the per site mutation rate parameter (θw) (Watterson 1975), were computed for each gene in each species using DnaSP v5 (Librado and Rozas 2009). Three neutrality tests based on the neutral frequency spectrum were performed for each locus in each species: Tajima’s D (Tajima 1989) and Fu and Li’s D and F using the alternate species as outgroup (Fu and Li 1993) as well as the MK test (McDonald and Kreitman 1991). For the \textit{atp5c1} gene, a between-species F\textit{ST} profile was generated using polymorphic sites, and between sites linkage disequilibrium (LD) was estimated using Haplovie v4.2 (Barrett et al. 2005). Phylogenetic relationships among haplotypes were then inferred using uncorrected P distances to reconstruct a neighbor-joining tree for the \textit{mt-atp6} gene and a NeighborNet network for the \textit{atp5c1} gene in the software SplitsTree4 (Huson and Bryant 2006). Finally, ancestral states of amino acids were estimated using the ML method in MEGA 5 (Tamura et al. 2011), at each node of an ML tree inferred from amino acid sequences of both Atlantic eel species and other available fish species as outgroups.

\textbf{Results}

\textbf{Sequence Assembly and SNP Discovery}

A total of 1,621,514 reads representing 472.8 Mb of sequences were de novo assembled into 22,014 contigs characterized by an average length of 459 bp and a mean number of 36.4 reads per contig. Among 55,602 putative SNPs which were identified in silico using the NQS algorithm, we retained 14,348 SNPs from 3,566 contigs (average of four SNPs per contig) after filtering for a minimal coverage of 8 × in both species. Using these parameters, we found a high level of polymorphism across the transcriptome of both Atlantic eel species, with an average number of one SNP every 114 bp.

\textbf{Identification of Genes with Diagnostic Mutations}

The distribution of allele frequency differences between species was skewed toward low values, with a median of 0.11 and 67.8% of the SNPs showing frequency differences below 0.20 (fig. 1). Only 231 highly divergent SNPs from 112 different contigs exhibited allele frequency differences above 0.95. Manual correction of local assembly and sequencing errors revealed a total of 94 annotated contigs showing at least one diagnostic mutation (mean = 2.9 species-diagnostic sites per contig). These were distributed among 87 nuclear genes (42 of which were fully sequenced) and 7 mitochondrial genes (4 of which were fully sequenced), for a mean number of 697 coding base pairs per contig (the complete list can be found in supplementary table S1, Supplementary Material online). Testing for overrepresented Gene Ontology terms in the subset of nuclear genes with diagnostic mutations compared with the reference list of the 3,566 genes retained for sufficient coverage revealed a highly significant enrichment in genes involved in the ATP biosynthetic process (GO:0006754, \textit{P} value = 0.000027, FDR-corrected \textit{P} value = 0.0024). ATP synthesis coupled with proton transport (GO:0015986), which is an ATP biosynthetic process, was also significantly over-represented (\textit{P} value = 0.000925, FDR-corrected \textit{P} value = 0.0151).

\textbf{Analyses of Molecular Evolution}

The mean number of synonymous substitutions per synonymous site (\textit{d}\textsubscript{s}) and the mean number of nonsynonymous substitutions per nonsynonymous site (\textit{d}\textsubscript{n}) calculated across all 87 nuclear diagnostic genes were 0.0048 and 0.0027.
Distribution of \( d_n, d_s \), and \( \omega \) across the 13 mitochondrial genes, ordered according to their relative positions in the mitochondrial genome of Atlantic eels.

respectively, for a mean nonsynonymous/synonymous rate ratio of \( \omega_n = 0.565 \). For the seven mitochondrial genes obtained by 454 pyrosequencing, the mean \( d_s \) was 8.35 times higher than for nuclear genes (\( d_s = 0.0405 \)) and the mean \( d_n \) was 2.75 times lower (\( d_n = 0.0010 \)), for a mean nonsynonymous/synonymous rate ratio of \( \omega_m = 0.025 \). The mean values of \( d_s \) and \( d_n \) estimated from all the 13 mitochondrial protein-coding genes were 0.1024 and 0.0021, respectively, for a mean \( \omega_m = 0.021 \) and a mitochondrial to nuclear synonymous substitution ratio of 21.3. The locus \( \text{mt-atp6} \) exhibited by far the highest nonsynonymous to synonymous rate ratio among the 13 mitochondrial genes (\( \omega = 0.126 \); fig. 2), being six times higher than the mean \( \omega_m = 0.021 \) ratio.

Contrasting indications of adaptive evolution were found across sequence data sets using each of the three methods to estimate \( \omega \) (Fay et al. 2001; Smith and Eyre-Walker 2002; Bierne and Eyre-Walker 2004 (supplementary table S2, Supplementary Material online)). Using polymorphic and divergent sites from the 87 nuclear diagnostic genes, estimated values of \( \omega_n \) ranged from 0.531 to 0.753 using polymorphism from \( A. \ rostrata \) and from 0.113 to 0.263 using polymorphism from \( A. \ anguilla \), while being only significant using the first species as ingroup (fig. 3). When controlling for the effect of SDMs segregating at low frequencies (\( \leq 20\% \)) in nuclear genes, significant values of \( \omega_n \) were obtained with each method using both species as ingroup but were still higher with \( A. \ rostrata \) (\( \omega_n = 0.822–0.957 \)) compared with \( A. \ anguilla \) (\( \omega_n = 0.440–0.660 \)).

Using the data set comprising seven mitochondrial genes (which does not include the \( \text{mt-atp6} \) gene), significantly negative values of \( \omega_m \) were inferred with the BEW method (Bierne and Eyre-Walker 2004) using polymorphism from each species (\( A. \ rostrata \); \( \omega_m = -6.985 \), \( A. \ anguilla \); \( \omega_m = -6.554 \); supplementary table S2, Supplementary Material online). However, these values became nonsignificant when low-frequency mutations were removed from the analyses (\( A. \ rostrata \); \( \omega_m = -2.571 \), \( A. \ anguilla \); \( \omega_m = -1.813 \)).

As expected in the presence of SDMs, within-species allele frequency spectra analyses revealed decreasing ratios of nonsynonymous to synonymous polymorphisms (\( p_n/p_s \) with increasing frequency classes for both nuclear and mitochondrial data sets in each species (fig. 4). The decrease in the \( p_n/p_s \) ratio was faster in \( A. \ rostrata \) for both data sets, but significant differences in \( p_n/p_s \) ratios between species were only observed with nuclear genes. The proportion of neutral nonsynonymous mutations, measured by the \( p_n/p_s \) ratio of common polymorphisms (>50%), was 0.11 in \( A. \ rostrata \) and 0.12 in \( A. \ anguilla \) based on the nuclear data sets and was not significantly different in the mitochondrial data sets (\( p_n/p_s = 0.08 \) and 0.06 for \( A. \ rostrata \) and \( A. \ anguilla \), respectively).

Analysis of Two Subunits of the ATP Synthase Cytonuclear Complex

Molecular diversity indices calculated from 20 \( \text{mt-atp6} \) sequences in each species were 2.5–4 times lower in \( A. \ rostrata \) compared with \( A. \ anguilla \) (table 1 and fig. 5a). Significant negative values of \( F \) and \( \Delta \) showed an excess of low-frequency mutations in \( A. \ anguilla \), indicating either a recent population expansion or directional selection. The MK test showed a significant excess of fixed amino acid–altering substitutions between species, consistent with positive selection (table 2). Among the six amino acid replacement substitutions found in the \( \text{mt-atp6} \) gene, two corresponded to a change that occurred in \( A. \ rostrata \) (I43T and M83T) and three in \( A. \ anguilla \) (the double-nucleotide mutation N52G and F105L), whereas the ancestral state could not be determined at the V195I substitution site.

Resequencing data from the most diverging region of the \( \text{atpSc} \) gene showed a 2.7-fold lower nucleotide diversity in \( A. \ anguilla \) compared with \( A. \ rostrata \) and significant negative values of Tajima’s \( D \) and \( F \) showed an excess of low-frequency mutations in \( A. \ anguilla \), indicating either a recent population expansion or directional selection. The MK test showed a significant excess of fixed amino acid–altering substitutions between species, consistent with positive selection (table 2). Among the six amino acid replacement substitutions found in the \( \text{mt-atp6} \) gene, two corresponded to a change that occurred in \( A. \ rostrata \) (I43T and M83T) and three in \( A. \ anguilla \) (the double-nucleotide mutation N52G and F105L), whereas the ancestral state could not be determined at the V195I substitution site.

Resequencing data from the most diverging region of the \( \text{atpSc} \) gene showed a 2.7-fold lower nucleotide diversity in \( A. \ anguilla \) compared with \( A. \ rostrata \) and significant negative values of Tajima’s \( D \) and \( F \) showed an excess of low-frequency mutations in \( A. \ anguilla \), indicating either a recent population expansion or directional selection. The MK test showed a significant excess of fixed amino acid–altering substitutions between species, consistent with positive selection (table 2). Among the six amino acid replacement substitutions found in the \( \text{mt-atp6} \) gene, two corresponded to a change that occurred in \( A. \ rostrata \) (I43T and M83T) and three in \( A. \ anguilla \) (the double-nucleotide mutation N52G and F105L), whereas the ancestral state could not be determined at the V195I substitution site.

![Fig. 2. Distribution of \( d_n, d_s \), and \( \omega \) across the 13 mitochondrial genes, ordered according to their relative positions in the mitochondrial genome of Atlantic eels.](image-url)
(I212V), whereas the ancestral state could not be determined at the A195V substitution site.

**Discussion**

**Evidence for Positive Selection at Nuclear Diagnostic Genes**

Analyzing the empirical distribution of allele frequency differences between species revealed that up to 2.6% of the transcribed genes (94/3,566) may present differentially fixed SNPs. This proportion is likely to be a slight overestimate of the true percentage of diagnostic SNPs across the transcriptome of Atlantic eels since our minimum base coverage threshold of 8× per species may not allow detecting rare variants that segregate in each genetic background. Moreover, because SNPs were detected from RNA-Seq data, gene specific variance in expression level among individuals or among alleles may contribute to bias in allele frequency estimates from pooled sample reads in *A. anguilla*. Nevertheless, our coverage threshold was reasonably high compared with similar studies of this type in nonmodel organisms (Galindo et al. 2010; Renaut et al. 2010; Barreto et al. 2011), which should minimize this problem. Overall, the distribution of interspecies allele frequency differences was consistent with the low-background level of neutral differentiation assessed with amplified fragment length

![Figure 3](image.png)

**Figure 3.** Estimates of α from the nuclear gene data set using (a) *Anguilla rostrata* and (b) *Anguilla anguilla* as ingroup. Three estimation methods were used: FWW (Fay et al. 2001), SEW (Smith and Eyre-Walker 2002), and BEW (Bierne and Eyre-Walker 2004), each using either all polymorphic sites (open squares) or after removing low- to moderate-frequency mutations (≤20%) segregating in each species (solid squares). Mean estimates are shown with their 95% confidence intervals.

![Figure 4](image.png)

**Figure 4.** Ratios of nonsynonymous to synonymous polymorphisms (pₙ/pₛ) within each species for low- (≤5%), moderate- (5–20%), common- (20–50), and high-frequency (>50%) classes separately calculated from (a) the nuclear and (b) mitochondrial gene data sets. Differences between pₙ/pₛ ratios were tested for statistical significance using the chi-square test. Significant differences between species for each frequency class are indicated above histograms in a and b and significant differences between frequency classes within each species appear in tables c and d (*Anguilla rostrata* above diagonal and *Anguilla anguilla* below diagonal). Aro: *A. rostrata*, Aan: *A. anguilla*, *P < 0.05, **P < 0.01, and ***P < 0.001.
polymorphism (Gagnaire et al. 2009) and microsatellite markers (Wirth and Bernatchez 2003) and with the view that most of the genome is still permeable to gene flow between species. Thus, even though some of the 87 nuclear genes may not be fully diagnostic because alleles from the alternate species possibly segregate at low frequencies in each species, these genes are among the most likely candidates for identifying genetic barriers involved in partial reproductive isolation since they exhibit little or no introgression.

In recently diverged species with large effective population sizes such as Atlantic eels (Wirth and Bernatchez 2003), most neutral alleles may not have had enough time to reach reciprocal monophyly since the onset of divergence. Thus, incomplete lineage sorting, in addition to introgression, likely contributes to the low proportion of reciprocally monophyletic loci (Funk and Omland 2003). In such a context, a stochastic steady state has probably not been reached for both neutral and adaptive mutations (Bierne and Eyre-Walker 2004) since advantageous alleles are the only one that may have had enough time to reach fixation. This stochastic steady state will take up to \(10N_e\) generations before being established after the species split (Nei and Li 1976). Therefore, \(\alpha\) will probably overestimate the true proportion of substitutions driven by positive selection during this period. This possibility was previously suggested in a comparative study of plants (Gossmann et al. 2010), where the highest values of \(\alpha\) were found between the most recently diverged species. Because of this potential bias, our estimates of \(\alpha\) may not be directly comparable to the rates of adaptive evolution estimated using well-separated species (Eyre-Walker 2006).

Although recent divergence between young species will not generate significant values of \(\alpha\) under neutrality, artificial evidence for positive selection may be obtained if there is an increase in population size and slightly deleterious amino acid polymorphisms (McDonald and Kreitman 1991; Eyre-Walker 2002). In Atlantic eels, however, this possibility is less likely given the demographic contraction recently experienced by both species (Wirth and Bernatchez 2003), which probably reduced the selective constraints on SDMs. Moreover, the hypothesis of a recent population expansion was not supported by the excess of amino acid polymorphisms relative to substitutions in mitochondrial genes (Nachman et al. 1996; Eyre-Walker 2002), as revealed by our negative estimates of \(\bar{x}_{mito}\). Differences in estimated values of \(\alpha\) between nuclear and mitochondrial data sets have previously been observed in other species and attributed to differences in rates of recombination combined with the presence of SDMs (Weinreich and Rand 2000). Our results also suggested that only a small proportion of nonsynonymous mutations may be effectively neutral in Atlantic eels (6–12%) and that nuclear and mitochondrial genes experience similar intensities of selective constraints. These observations are thus consistent with the finding of lower values of \(\alpha\) in the nonrecombining mitochondrial DNA (mtDNA; Weinreich and Rand 2000) due to the segregation of SDMs.

### Table 1. Molecular Diversity Indices and Results of Neutrality Tests for the mt-atp6 and the atpSc1 Genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>Length (bp)</th>
<th>n</th>
<th>S</th>
<th>h</th>
<th>k</th>
<th>(\pi)</th>
<th>(\theta_N)</th>
<th>Tajima’s D</th>
<th>Fu and Li’s D</th>
<th>Fu and Li’s F</th>
</tr>
</thead>
<tbody>
<tr>
<td>mt-atp6</td>
<td>Aro</td>
<td>684</td>
<td>20</td>
<td>7</td>
<td>6</td>
<td>1.974</td>
<td>0.0029</td>
<td>0.0029</td>
<td>-0.0429</td>
<td>-1.0136</td>
<td>-0.9429</td>
</tr>
<tr>
<td></td>
<td>Aan</td>
<td>684</td>
<td>20</td>
<td>28</td>
<td>18</td>
<td>4.9842</td>
<td>0.0073</td>
<td>0.0124</td>
<td>-1.6093</td>
<td>-2.8623**</td>
<td>-2.9853**</td>
</tr>
<tr>
<td>atpSc1</td>
<td>Aro</td>
<td>1,627</td>
<td>40</td>
<td>59</td>
<td>36</td>
<td>11.5410</td>
<td>0.0081</td>
<td>0.0098</td>
<td>-0.6032</td>
<td>-1.7859</td>
<td>-1.6254</td>
</tr>
<tr>
<td></td>
<td>Aan</td>
<td>1,627</td>
<td>40</td>
<td>38</td>
<td>31</td>
<td>4.2372</td>
<td>0.0030</td>
<td>0.0063</td>
<td>-1.8419*</td>
<td>-2.1119*</td>
<td>-2.1119*</td>
</tr>
</tbody>
</table>

Note.—Aro: Anguilla rostrata, Aan: Anguilla anguilla; n: number of sequences, S: number of segregating sites, h: number of haplotypes, k: average number of nucleotide differences, \(\pi\): nucleotide diversity, and \(\theta_N\): Watterson’s estimator of the per site mutation rate parameter.

* \(P < 0.05\), ** \(P < 0.01\).

### Table 2. MK Test for the mt-atp6 and atpSc1 Genes.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Substitution/Polymorphism</th>
<th>Replacement</th>
<th>Synonymous</th>
</tr>
</thead>
<tbody>
<tr>
<td>mt-atp6</td>
<td>Fixed differences between species</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>Polymorphic sites within species</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>One-tailed Fisher’s exact test: (P) value = 0.030</td>
<td></td>
<td></td>
</tr>
<tr>
<td>atpSc1 (Exons 6–8)</td>
<td>Fixed differences between alleles</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Polymorphic sites within alleles</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>One-tailed Fisher’s exact test: (P) value = 0.049</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—The MK test was performed between species for the mt-atp6 gene and between European and American alleles for the coding region comprising exons 6–8 of the atpSc1 gene.

---

Fig. 5. Neighbor-joining tree for the entire mt-atp6 gene (a) and neighbor network for the intron 5 to exon 8 region of the atpSc1 gene (b). Anguilla rostrata haplotypes are represented by open circles and Anguilla anguilla by black circles. Unidirectional introgression at the atpSc1 locus is illustrated by one open circle in the A. anguilla part of the neighbor network showing introgression of a European allele into the American eel background and by six American haplotypes with intermediate positions in the network showing recombination of American with European alleles around the G211A-V212I double mutation.
We found that, on average, nonsynonymous mutations segregate at lower frequencies than synonymous polymorphisms in both species, which confirms that SDMs segregate at low frequencies in the cytoplasmic and nuclear genomes of each species (Fay et al. 2001, 2002; Zhang and Li 2005; Parsch et al. 2008). Since we made the assumption that nonsynonymous polymorphisms are neutral to perform MK-based estimations of $\alpha$, low-frequency SDMs will probably make MK-related tests more conservative due the recent contraction in effective population size experienced by both species (Bierne and Eyre-Walker 2004). This prediction is consistent with the higher values of $\alpha$ systematically obtained when excluding polymorphisms below 20% to control for the effects of negative selection acting on SDMs, as previously observed in studies on human and Drosophila (Fay et al. 2001, 2002; Bierne and Eyre-Walker 2004). Although removing low-frequency classes from the analysis may still underestimate the rate of adaptive evolution (Charlesworth and Eyre-Walker 2008), this treatment resulted in significant positive values of $\alpha_{\text{nucl}}$ in each species, with each of the three MK-related tests. Thus, significant excess of nonsynonymous to synonymous substitutions relative to polymorphisms provides evidence for positive selection acting on at least some nuclear diagnostic genes included in our analysis. In parallel, we did not find any evidence for positive selection acting on our subset of seven mitochondrial genes, even when low- and moderate-frequency mutations were removed from the analysis. Instead, our results suggested that those genes (which do not include the mt-atp6 gene) mostly evolve under strong purifying selection.

**Indications for the Existence of a Cytonuclear Incompatibility**

Contrasting with the fact that most of the nuclear genome is still permeable to interspecific gene exchange, the absence of mitochondrial introgression suggests the existence of asymmetrical crossing barriers with respect to sex or postzygotic isolation mechanisms involving one or several mitochondrial genes. Since both American and European mitochondrial haplotypes have been found in Icelandic hybrids (Avise et al. 1990; Frankowski and Bastrop 2010), sex-based directionality in interspecific crosses does not seem to explain the barrier to mitochondrial gene flow. On the other hand, the significant enrichment of genes involved in the ATP synthesis coupled with proton transport process among nuclear diagnostic genes revealed some candidate loci potentially responsible for cytonuclear hybrid dysfunction. Of the 12 nuclear-encoded proteins that compose the ATP synthase complex, 5 were found to be divergent in our transcriptome scan approach. Moreover, among the two mtDNA-encoded proteins in the complex, the ratio of nonsynonymous to synonymous substitutions displayed by the mt-atp6 gene was six times higher than the mean $\alpha$ value of the 13 mitochondrial genes, indicating either positive selection or relaxed purifying selection. Although demonstrating coadaptation between mtDNA- and nuclear-encoded proteins would require cytonuclear experiments or introgression analyses using natural backcrosses (Blier et al. 2001; Rand et al. 2004), resequencing the mt-atp6 gene and the most divergent nuclear gene in the ATP synthase complex (atp5c1) enabled the use of indirect approaches based on population genetic neutrality tests. Significant excesses of replacement/synonymous substitutions relative to polymorphisms strongly suggested that both the mt-atp6 and the atp5c1 genes have undergone divergent selection between species. Therefore, even if we found that SDMs segregate in the mitochondrial genome of each species, the fixation of mildly deleterious mutations due to relaxed purifying selection seems insufficient to explain the number of nonsynonymous substitutions at the mt-atp6 locus. Moreover, a scenario whereby positive selection has recently driven amino acid replacements at this gene in both species was also supported by a strikingly low level of nucleotide diversity in A. rostrata and a significant excess of low-frequency polymorphisms in A. anguilla.

Species with high effective population sizes may experience recurrent selective sweeps (Bazin et al. 2006) provoking transient reductions of nucleotide diversity in the nonrecombining mitochondrial genome. Contrary to genetic drift...
predictions, this genetic draft effect makes the fixation rate of SDMs increasing with the population size (Gillespie 2001). The fixation of SMDs due to genetic hitchhiking on the mitochondrial chromosome may subsequently elicit the fixation of compensatory mutations within the same gene or within nuclear genes coding for interacting proteins (Rand et al. 2004; Oliveira et al. 2008). Evidence for positive selection at the nuclear gene atp5c1, encoding for a protein that interacts with the ATP synthase F0 subunit 6, may be seen as an indication of such compensatory coadaptation between two interacting genes. Furthermore, species-diagnostic mutations that were found in four other nuclear genes encoding subunits of the ATP synthase raise the possibility that compensatory mutations have also been selected in other nuclear genes implicated in the oxidative phosphorylation complex V. An exhaustive analysis of polymorphism and divergence of all the genes involved in the ATP synthase complex would accordingly provide a better understanding of this substitution pattern (Gibson et al. 2010). However, it appears likely from our results that positive selection has caused the fixation of different amino acid mutations in each species, at least for two ATP synthase subunits, one being encoded by the cytoplasmic and the other by the nuclear genome. Because these putatively species-specific associations of coadapted alleles will be disrupted in F2 hybrids, maladaptive combinations in recombined cytonuclear complexes may cause genetic barriers to interspecific gene flow and thus contribute to reproductive isolation (Burton et al. 2006).

Dobzhansky–Muller incompatibilities involving the interaction of mitochondrial and nuclear genes have been found in several organisms including Tigriopus (Willett and Burton 2001; Burton et al. 2006), Drosophila (Sackton et al. 2003; Montooth et al. 2010), Nasonia (Ellison et al. 2008; Niehuis et al. 2008; Oliveira et al. 2008), and Saccharomyces (Lee et al. 2008; Chou et al. 2010). In those species, selection for compensatory mutations in response to the fixation of SMDs by drift or genetic hitchhiking in the mitochondrial genome has been proposed to explain the evolution of postzygotic isolation, which may be accelerated by a higher mitochondrial mutation rate (Oliveira et al. 2008). In Atlantic eels, the mitochondrial to nuclear synonymous substitution ratio of 8.4–21.3 lies between the 2.75 value found in Drosophila (Montooth et al. 2009) and the ratios above 25 estimated in Nasonia (Oliveira et al. 2008) and Tigriopus (Willett and Burton 2004). While this disparity could be explained by differences in mutation rates or efficiency of purifying selection, the lower rate observed in Drosophila also coincides with a lack of fixed cytonuclear incompatibilities between fly species (Montooth et al. 2010). These results suggest that a higher mitochondrial substitution rate could make the evolution of cytonuclear incompatibility more likely.

Deleterious epistatic interactions between nuclear-encoded proteins are also known to contribute to reproductive isolation in several model species (Presgraves 2010). Such nuclear–nuclear incompatibilities may be present in Atlantic eels as well, and their accumulation may participate to the build-up of reproductive isolation in the presence of gene flow (Barton and de Cara 2009). Although we did not identify all the genes which are resistant to introgression, the new list of nuclear diagnostic genes revealed in our study sets the stage for further investigations of their possible roles in hybrid breakdown.

Nevertheless, the disproportionate mapping of diagnostic markers to the ATP synthase complex may be indicative of the importance of mitochondrial–nuclear conflicts in postzygotic isolation between Atlantic eels and may explain why asymmetrical patterns of introgression have been observed in previous studies (Avise et al. 1990; Albert et al. 2006; Gagnaire et al. 2009). Surprisingly, the direction of introgression at the atp5c1 gene was opposite to that observed for most nuclear markers, which preferentially introgress from the American to the European background. Several explanations may account for this pattern, such as the number of genes involved in the cytonuclear incompatibility, the relative timing of the mutations, their fitness effects, and the recessivity of incompatible interactions (Orr 1995). Asymmetric introgression in the flanking regions of the G211A-V212I double mutation in the exon 7 of atp5c1 resulted in a rapid decay of genetic differentiation and LD on each side of the candidate site. This result, indicative of a relatively short region of genomic differentiation spanning only a few exons and introns around the candidate functional mutation, is consistent with the observation that differentiation does not extend beyond a few kilobases in some model organisms (Turner et al. 2008, 2010; Kolaczkowski et al. 2011). With the increasing sequencing power offered by NGS methods, further screening of the chromosomal signature at genetic incompatibility loci will enable evaluating the generality of this pattern in a context of early divergence between partially reproductively isolated species.

In conclusion, we showed that adaptive protein evolution has occurred at several of the few genes showing between-species divergence despite pervasive introgression. This finding is consistent with the action of positive selection in facilitating the establishment and maintenance of genetic barriers to gene flow during the early stages of speciation. Contrasting with the prevailing view that candidate loci detected by genome scans are attributable to ecological adaptation, our transcriptome scan performed in Atlantic eels illustrates the potential of such methods to identify intrinsic factors (Bierne et al. 2011) and ultimately speciation genes contributing to reproductive isolation.

**Supplementary Material**

Supplementary tables S1 and S2 and figure S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We are grateful to the Associate Editor Michael Nachman as well as three anonymous referees for their constructive and helpful comments on the manuscript. We also thank M.M. Hansen, L. Zane and G.E. Maes from the Anguilla
Genomics Consortium. This research was supported by the Natural Science and Engineering Research Council of Canada (NSERC) and a Canadian Research Chair in Genomics and Conservation of Aquatic Resources to L.B. P.A.G. was supported by a Government of Canada Post-Doctoral Fellowship.

References


_Hereditas_ 56:337–349.


