A 39-kb Sequence Around a Blackbird Mhc Class II Gene: Ghost of Selection Past and Songbird Genome Architecture

Scott V. Edwards,* Joe Gasper,* Daniel Garrigan,*1 Duane Martindale,† and Ben F. Koop†

*Department of Zoology, University of Washington; and †Center for Environmental Health, Department of Biology, University of Victoria, British Columbia, Canada

To gain an understanding of the evolution and genomic context of avian major histocompatibility complex (Mhc) genes, we sequenced a 38.8-kb Mhc-bearing cosmid insert from a red-winged blackbird (Agelaius phoeniceus). The DNA sequence, the longest yet retrieved from a bird other than a chicken, provides a detailed view of the process of gene duplication, divergence, and degeneration (“birth and death”) in the avian Mhc, as well as a glimpse into major noncoding features of a songbird genome. The peptide-binding region (PBR) of the single Mhc class II B gene in this region, Agph-DAB2, is almost devoid of polymorphism, and a still-segregating single-base-pair deletion and other features suggest that it is nonfunctional. Agph-DAB2 is estimated to have diverged about 40 MYA from a previously characterized and highly polymorphic blackbird Mhc gene, Aph-DAB1, and is therefore younger than most mammalian Mhc paralogs and arose relatively late in avian evolution. Despite its nonfunctionality, Agph-DAB2 shows very high levels of nonsynonymous divergence from Aph-DAB1 and from reconstructed ancestral sequences in antigen-binding PBR codons—a strong indication of a period of adaptive divergence preceding loss of function. We also found that the region sequenced contains very few other unambiguous genes, a partial Mhc-class II gene fragment, and a paucity of simple-sequence and other repeats. Thus, this sequence exhibits some of the genomic streamlining expected for avian as compared with mammalian genomes, but is not as densely packed with functional genes as is the chicken Mhc.

Introduction

The genomes of various vertebrate groups are characterized by differences in size, gene and repeat density, and isochore composition. We expect these features to be reflected in the genomic structure of multigene families of these groups. For example, avian genomes are ~50% smaller than those of mammals (Tiersch and Wachtel 1991), and chicken genomes are depauperate in simple sequence repeats (Pimmer et al. 1997), have higher GC% contents (Bernardi, Hughes, and Mouchiroud 1997), higher gene densities (McQueen et al. 1996), and smaller introns (Hughes and Hughes 1995) than do those of mammals. These genomic differences between birds and mammals have been suggested to have their origin in lineage-specific selection for small cell and genome size imposed by flight and its associated metabolic and behavioral demands (Tiersch and Wachtel 1991; Hughes 1995) or possibly other unknown selective agents.

The major histocompatibility complex (Mhc) of vertebrates, a region containing the most polymorphic genes in the vertebrate genome (Edwards and Hedrick 1998), many of which have functions in defense against pathogens, appears to reflect some of these trends. For example, the chicken Mhc (~100 kb) is orders of magnitude smaller than the Mhc of mice and humans (~4 Mb) (Trowsdale 1995). It has long been known that chicken Mhc genes possess much smaller introns than those of mammalian Mhc genes (Kaufman, Salamonsen, and Flajnik 1991), and it was recently shown that, at one gene per 5 kb, the chicken Mhc (B complex) is much more gene dense than the class I or II regions of mammals (Kaufman et al. 1999b). These differences suggest that the chicken Mhc may have responded to the same selective pressures as the rest of the avian genome.

The smaller number of Mhc genes in the “minimal essential” chicken Mhc is thought to focus parasite-mediated selection adaptively on a few target genes, thereby resulting in associations between specific haplotypes and disease resistance that are stronger than those observed in mammals (Kaufman 1995). In addition, the specific organization and tight linkage of genes in the chicken Mhc has been suggested to facilitate coevolution of functionally associated protein products, such as Mhc class I and peptide transporters (TAP) (Kaufman et al. 1999a). However, we know little about the genomic organization of Mhcs in birds other than the chicken with which to test the generality of these structural features. Coding sequences of Mhc genes in songbirds and game birds suggest that the long-term pattern of class II gene evolution in birds is characterized by higher rates of concerted evolution, or more recent postspeciation duplications of genes, than are found in mammals (Edwards, Wakedland, and Potts 1995; Edwards et al. 1999; Wittzel et al. 1999). However, some songbirds exhibit a greater complexity of class II genes on Southern blots than do chickens (Edwards, Nusser, and Gasper 2000), and a recent molecular analysis of class I genes in the great reed warbler (Acrocephalus arundinaceus) suggested a much greater number of expressed class I genes in this songbird than in chickens (Westerdahl, Wittzell, and von Schantz 1999). Thus, it is not clear the extent to which evolutionary trends and genomic organization of chicken Mhc genes will represent those of songbirds and other avian lineages.

Understanding in detail the long-term evolution of the Mhc in birds requires appropriate phylogenetic sam-
pling at both the root and the tips of the avian tree (Edwards et al. 1999). It has recently been proposed on the basis of complete mitochondrial genome sequences that perching birds (Passeriformes) may represent a basal lineage in birds, perhaps the sister group to all other birds (Härlid and Arnason 1999; but see Groth and Barrowclough 1999; van Tuinen, Sibley, and Hedges 2000). Thus, pending clarification of the phylogenetic placement of perching birds, it would be useful to gain insight into \textit{Mhc} structure in this lineage. We recently characterized \textit{Mhc} class II B genes at the genomic level in two songbirds, red-winged blackbirds (\textit{Agelaius phoeniceus}) and house finches (\textit{Carpodacus mexicanus}) (Edwards, Gasper, and Stone 1998; Hess et al. 2000), both of which have served as models in ecology and evolution research (Beletsky 1996). Further knowledge of the \textit{Mhc}s of these species would also be useful for understanding the genetic basis of disease resistance and mate choice in natural populations. We recently used shotgun sequencing to study the larger genomic context and evolution of \textit{Mhc} genes in house finches (Hess et al. 2000). Here, we use a similar strategy to characterize further \textit{Mhc} genes in blackbirds.

**Materials and Methods**

**Cosmid Subcloning, Sequencing, and Contig Analysis**

Although the molecular methods we used in this paper are not novel from the standpoint of model organism genomics, we describe them in some detail because they may be new to some readers of this journal. We chose to sequence a red-winged blackbird cosmid clone (RWcos10) that was isolated from the same library (from a female blackbird) that yielded a previously sequenced blackbird class II gene, \textit{Agph-DAB1}, but had a different restriction map and \textit{Mhc}-probed Southern blot profile (Edwards, Gasper, and Stone 1998). Preliminary sequencing confirmed that the \textit{Mhc} gene on \textit{Rwcos10} was a distinct locus from \textit{Agph-DAB1}. Details of construction and screening of the cosmid library, generated via partial digestion of blackbird DNA and ligation into \textit{sCos}-1 vector, are provided elsewhere (Edwards, Gasper, and Stone 1998; Edwards, Nusser, and Gasper 2000). To sequence this cosmid, we used the same techniques as those used in human genomics (Rowen and Koop 1994). Briefly, the entire insert and vector were sonicated using a cup horn sonicator. The ends of these fragments, and a band corresponding to 2.5–4-kb fragments was excised from the gel. The sonicated DNA fragments were agarose electrophoresed, and a band corresponding to 2.5–4-kb fragments was excised from the gel. The ends of these fragments, which were ragged due to the sonication, were made blunt by T4 DNA polymerase and subcloned into M13. Several hundred M13 plasmids were grown and prepared for sequencing using a 96-well plate format (Huang 1994). Prior to targeted sequencing of specific subclones to complete contig assembly, we sequenced 928 randomly selected clones on an ABI373A DNA sequencer using dye-terminator chemistry and a modified M13 forward primer that eliminated recovery of plasmid sequence.

New chromatograms were generated from the raw sequence data using the base-calling program PHRED (Ewing and Green 1998; Ewing et al. 1998). The sequence reads were assembled into contigs using both sequence overlap and sequence quality information by the program PHRAP (P. Green, unpublished; http://bozeman.mbt.washington.edu/phrap.docs/phrap.html). The resulting contigs and chromatogram data were visualized using CONSED (Gordon, Abajian, and Green 1998), which was also used to develop sequence closure strategies. We created larger contigs via extension of subclone sequences. Sequences from this study have been deposited in the GenBank database under accession numbers AF170972 (cosmid) and AF181836–AF181841 (\textit{Agph-DAB2} alleles).

**Gene Finding and Sequence Analysis**

The program SeqHelp (Lee, Lynch, and King 1998) was used to identify coding regions and putative exons using the internal module Genefinder (C. Wilson and P. Green, unpublished). Genefinder conducts BLAST searches for 6-kb segments of the input sequence and therefore provides an opportunity to find all potentially similar sequences in the GenBank database. We also used a modified version of the program GeneMark (Lukashin and Borodovsky 1998) (http://dixie.biology.gatech.edu/GeneMark/eukhmms.cgi) to identify potential open reading frames and exons. GeneMark uses a hidden Markov model to recognize statistical patterns in DNA sequences based on rules for primary structures of coding and noncoding regions, such as spacing of splice signals and start and stop codons. The algorithm utilizes rules in a matrix form determined from empirical examination of coding and noncoding regions of particular organisms; we used the matrix for chickens, \textit{Gallus gallus}. To identify simple sequence repeats (SSRs) and transposable elements, we used an internal module in SeqHelp called RepeatMasker (http://www.genome.washington.edu/UWGC/analysistools/repeatmask.htm), as well as a program called Sputnik (C. Abajian, unpublished). We found the latter to be more effective at finding very short SSRs. The criteria used by the Sputnik module for identifying SSRs was a repeat unit length of 2–5 and a minimum match score of 8 (1 point = single-base-pair match; −6 points = mismatch, insertion, or deletion). By these criteria, a perfect three-repeat SSR of unit length 2 would not be detected, whereas a perfect two-repeat SSR of unit length 4 would.

**PCR Survey and Polymorphism Analysis**

To examine genetic diversity in the peptide-binding region (PBR), we conducted a survey of polymorphism in the PBR-encoding second exon of the \textit{Mhc} gene found on RWcos10, \textit{Agph-DAB2}. We used eight birds from Kentucky, Florida, and New York, from which genomic DNA was isolated from blood by standard phenol-chloroform extraction methods. We designed two PCR primers that were targeted to flanking introns 1 and 2 and amplified a 395-bp segment spanning the entire second exon of \textit{Agph-DAB2} (\textit{rwcos10intf.2}: CCTGACCAGCCTCATGGA; \textit{rwcos10int2r.1}: ACGCTCTGCTCCGCGCT). We ligated...
these PCR products into Bluescript vector and sequenced five clones per individual. Sequences were aligned manually (Gilbert 1995). Two measures of polymorphism, the average number of pairwise differences per site (π) and a coalescent estimate of $\Theta = 4N_e \mu$ (where $N_e$ is the effective population size and $\mu$ is the mutation rate) with no population growth, were calculated from all aligned AgphDAB2 sequences using the programs DnaSP (Rozas and Rozas 1997) and Fluctuate (Kuhner, Yamato, and Felsenstein 1998), respectively. Both measures are in units of substitutions per site per generation. We tested the neutral-mutation hypothesis for these sequences using Tajima’s (1989) $D$ statistic. The age of particular classes of Agph-DAB2 alleles was estimated with a maximum-likelihood (ML) and Monte Carlo method (Slatkin and Rannala 1997) using a value of $N_e$ extrapolated from mtDNA data for red-winged blackbirds (Ball et al. 1988). For interlocus comparisons of class II B genes, the numbers of synonymous and nonsynonymous substitutions per site were calculated by Jukes-Cantor (Nei and Gojobori 1986) and ML (Goldman and Yang 1994) methods. Total divergence in coding and noncoding regions was estimated by the method of Tamura and Nei (1993). Reconstruction of inferred ancestral peptide-binding regions was conducted with the ML method of Yang, Kumar, and Nei (1995) using the modified PAM matrix of Jones, Taylor, and Thornton (1992). Relative-rate tests were conducted using two-cluster (Takezaki, Rzhetsky, and Nei 1995) and ML (Goldman and Yang 1994) methods.

**Results**

Features of the Blackbird Sequence

The insert of the blackbird Mhc-bearing cosmid clone Rwcos10 was 38,785 bp long (fig. 1). In addition to containing a full-length Mhc class II B gene (figs. 1B and 2), which we designate Agph-DAB2, SeqHelp identified two other protein-coding regions with convincing similarities to genes in the GenBank database: a fragment of an Mhc class II B (DAB) gene including sequences downstream but not upstream of exon 3 (figs. 1B and 2) and a zinc-finger domain of the C2H2 type (fig. 1B and table 1; Becker et al. 1995). However, both of these gene regions appear to be shorter than putative homologs, contain in-frame stop codons, and are likely pseudogenes.
Other regions identified by Seqpup as having significant amino acid similarity to known genes were four segments whose predicted amino acid sequences resembled parts of collagen genes (table 1 and fig. 1B). Putative collagen-containing regions varied in length from 12 to 34 amino acids, with similarity ranging from 35% to 50%. These segments were similar to conserved regions of collagen genes, and it is not clear at this time whether similarity to adjacent segments of collagen genes is present in cosmid sequence. Three collagen fragments occurred very close together and may represent a smaller number of genes or gene fragments (fig. 1B). Finally, SeqHelp found similarity of an inferred amino acid sequence from the cosmid to a retinoic acid receptor β gene (RXRB; table 1), which encodes an MHC class I regulatory element (Nagata et al. 1994).

At the DNA level, SeqHelp identified two regions (10714–10765 and 12224–12252) that exhibited substantial similarity (81% and 79%, respectively) to an mRNA for human neurotrypsin (fig. 1). In addition, a total of five intriguing but very short (20–35 bp) regions in the blackbird sequence bore some similarity (72%–95%) to noncoding regions in the chicken MHC. These regions did not occur in regions immediately upstream of genes that could suggest that regulatory or other sequences and as such may be spurious. GeneMark identified a total of 42 putative exons falling into 8 putative genes (fig. 1C). Three of these predicted exons corresponded exactly to exons 3 and 4 of Agph-DAB2 and to exon 5 of the DAB fragment (fig. 1C). An additional two predicted exons corresponded closely to the zinc-finger domain and a short but unconvincing segment of DNA or amino acid sequence similarity predicted by SeqHelp (fig. 1C). A complete description of putative matches to previously characterized sequences will be published elsewhere.

In sliding windows of 100 bp in length, the GC content of the 39-kb segment varies from 31% to 78%, with sustained peaks (>55%) in and upstream of both Agph-DAB2 and the DAB fragment. SeqHelp identified two clusters of potential CpG islands, which are often good indicators of genes in avian genomes (McQueen et al. 1996), that coincide with the two high GC peaks (fig. 1A). The exons predicted by GeneMark also fall largely in regions of elevated GC content. The cosmid insert also contained a total of 10 simple sequence repeats under liberal inclusion criteria (fig. 1D). However, three TA-rich SSRs fall immediately adjacent to one another and are clearly part of a single complex SSR, bringing the total number to 8. Additionally, only two of these, (CT)_{13} and (GGGAT)_{19}, were long enough to be polymorphic; the (CT)_{12} repeat is likely of borderline length with regard to polymorphism, and most of the others are far too short to expect polymorphism. The (CT)_{12} repeat occurred at the 3' end of intron 2 in the same position as a (CT), repeat in Agph-DAB1 and a TC-rich region in chicken class II B genes. Long pyrimidine tracts are frequently found in the 3' ends of introns (J. Kaufman, personal communication), and thus even this microsatellite is not surprising. In addition, RepeatMasker identified three putative transposable elements. One of these consisted of an envelope protein fragment that was clearly alignable to endogenous retroviruses (ERVs) of the human ERV-C type (Doolittle and Feng 1992), representatives of which are found in abundance in the human class I region (Kulski et al. 1999); however, two putative L1M-type LINE sequences (fig. 1E), which are common in the human class II region (Beck and Trowsdale 1999), were not confirmed by subsequent BLAST searches and are therefore tentatively considered false positives. A dot-plot analysis of the entire sequence to itself revealed no major repeat structures other than the homology of the Agph-DAB2 gene and the DAB fragment (not shown).

Structure and Polymorphism of Agph-DAB2

Analysis of Agph-DAB2 revealed it to be the longest avian Mhc gene characterized to date, 3,559 bp long from start to stop codon, including the five introns, which occurred in positions similar to those of chicken class II B genes (fig. 2). Agph-DAB2 is thus over three times as long as a typical chicken class II B gene and nearly 50% longer than Agph-DAB1. Agph-DAB2 possesses a poly-A site 147 bp downstream of the stop codon. Alignment of the DNA sequences of Agph-DAB1 and Agph-DAB2 showed that the cosmid clone sequence possessed a single-base-pair deletion 87 bp into exon 2 (fig. 2). This region of the gene was covered by six subclones of high quality sequence, indicating that it was not a sequencing error but leaving open the possibility that it was a cosm id cloning artifact. We therefore examined the sequence of Agph-DAB2 in this region as amplified directly from blackbird genomic DNA. A survey of eight birds (five clones per bird) revealed that the single-base-pair deletion was present in 3 of the 16 sampled chromosomes (~19%) and that the bird from which the library was made was heterozygous for the
deletion. Alignment of the inferred amino acid sequence of Agph-DAB2 consisting of exons 1, 3, and 4–6 and a nondeleted copy of exon 2 showed that the gene potentially encoded a full-length Mhc product of 261 amino acids, including three amino acid deletions (two in the leader peptide and one in exon 4) relative to a chicken BLBII gene (fig. 3). Of 19 residues deemed conservative and important for Mhc class II function (Kaufman, Salomonsen, and Flajnik 1994), Agph-DAB2 possesses 16, with aberrant residues at three exon 2 sites. In contrast, Agph-DAB1 possesses all 19 residues (fig. 3). Thus, although Agph-DAB1 and Agph-DAB2 exhibit a high level of conservation in exons, particularly those other than exon 2 (figs. 3 and 4), it is possible that even the alleles without the deletion are nonfunctional.

In addition to the indel polymorphism in exon 2, there were only six segregating sites in exon 2, yielding estimates of nucleotide (nonindel) diversity of $\pi = 0.0035$ and $\Theta = 0.0078$. These polymorphisms occurred in positions 108, 125, 140, 250, 254, and 268 of the exon and occurred twice in each of the three codon positions. The variation suggested a slight excess of rare mutations and a deviation from neutral evolution by Tajima’s test ($D = -1.624$, $P < 0.02$; significance tested using Monte Carlo simulation). To estimate the age of the deletion, we used information on $N_e$, the frequency of the deletion mutation, the amount of polymorphism of alleles belonging to the mutant class, and estimates of the mutation rate (Slatkin and Rannala 1997). We used two mutation rates: one derived from the silent rate of alleles belonging to the mutant class, and estimates of the deletion mutation, the amount of polymorphism in primate Mhc genes ($10^{-9}$ substitutions per site per year; Satta et al. 1993), and another of $3.84 \times 10^{-8}$, estimated as $\mu = \Theta/4N_e$, where $N_e = 50,000$. We extrapolated this value for $N_e$ from estimates of female $N_e$ from mtDNA data (Ball et al. 1988; see Garrigan and Edwards 1999 for details). There was no variability among the three alleles with the deletion, suggesting a relatively recent origin. The deletion was estimated to have occurred $\sim 830$ generations ago (about $\sim 1,000$ years, assuming a generation time of 1.3 years), with 95% confidence limits of $\sim 390$–$\sim 1,430$ generations ($\sim 500$–$\sim 1,860$ years). The ML estimate decreased to 520 generations ($\sim 680$ years) with an $N_e$ value of 10,000 and increased to 970 generations ($\sim 1,260$ years) with an $N_e$ value of 100,000.

Origin and Divergence Agph-DAB2

A comparison of Agph-DAB1 and Agph-DAB2 nucleotide sequences indicated that all exons, introns, and non-coding upstream regions were alignable except for intron 2, which was 571 bp longer in Agph-DAB2 (figs. 2 and 4). Silent divergence between the blackbird genes is significantly heterogeneous; in particular, silent divergence between Agph-DAB1 and Agph-DAB2 at the 5’ end of introns 1 and in exon 2 appear markedly higher than in other regions (fig. 4). The rank order of silent divergence for different regions (exon 2 > intron 1 > intron 3 > exon 3 > exon4/intron 5 > 5’ UT; see fig. 4 caption) and the observation that intron 2 is unalignable suggest that the interlocus divergence is in part a function of physical distance of a region from exon 2.

The phylogenetic relationships of Agph-DAB1 and Agph-DAB2 (fig. 5) are consistent with earlier analyses of expressed blackbird class II B sequences that revealed very high similarity among the then-known blackbird Mhc sequences and clustering of Mhc sequences in a species-specific manner (Edwards, Wakeland, and Potts 1995). Additionally, phylogenetic analysis of Agph-DAB1 and Agph-DAB2 and the DAB fragment based on 680 sites from intron 3 to exon 6 (not shown) suggests a bushlike, nearly simultaneous divergence of the three
genes from one another. We compared the divergence between Agph-DAB1 and Agph-DAB2 with that between human Mhc-DRB genes, whose time of origin has been extensively studied (Satta, Mayer, and Klein 1996). Using both two-cluster and ML methods, we conducted extensive molecular-clock analyses on a database of 44 mammalian and avian exon 1 and 3 sequences to test the hypothesis that blackbird and other songbird class II B genes are evolving at the same rate as human class II B genes (fig. 5). We consistently found that total rates in songbird exons exceeded those in chickens ($\delta = 0.10, Z = 2.12; P < 0.05$) but did not differ significantly from the rate among human DRB genes. In addition, we confirmed that silent transversional changes in songbird and human class II B genes were evolving at the same rate ($\delta = 0.16, Z = 0.406; P > 0.6$; the lintre program would not permit a test of total silent rates using our particular data set, perhaps because of invalid distances in some comparisons). Thus, these analyses indicate that two rate measures show similarity among songbirds and humans, making it likely that total silent rates in these two groups are also similar. We therefore employed absolute substitution rates for humans to the blackbird genes. If we assume a silent substitution rate of $10^{-9}$ substitutions per site per year, as has been estimated for introns and exons of human DRB genes (Satta et al. 1993; Satta, Mayer,
Fig. 4.—Silent divergence between blackbird class II B genes Agph-DAB1 and Agph-DAB2. The sliding window was 100 bp long, with a 10-bp offset length. The double slashes in the x-axis and the schematic gene diagram below indicate removal of ambiguously aligned sites from the full length of the intron. The dashed line above intron 2 also designates ambiguously aligned sites. The total silent divergence (Jukes-Cantor method) between the two genes for each region is as follows: 5'UT, 0.0220 ± 0.007; exon 1, 0.000 ± 0.000; intron 1, 0.0925 ± 0.017; exon 2, 0.2107 ± 0.068; exon 3, 0.0756 ± 0.0337; intron 3, 0.0460 ± 0.010; exon 4/intron4/exon 5, 0.0345 ± 0.008.

and Klein 1996), these data suggest that exon 3 of Agph-DAB1 and Agph-DAB2 (dS = 0.071) diverged approximately 35 MYA. The observation that the silent divergence of Agph-DAB1 and Agph-DAB2 is about 70% of that for DRB1 and pseudogene DRB7 (dS = 0.101), which diverged around 54 MYA (Satta, Mayer, and Klein 1996), suggests that the divergence of Agph-DAB1 and Agph-DAB2 occurred approximately 40 MYA. Silent divergences between Agph-DAB1 and Agph-DAB2 for various exons and introns are less than those for the corresponding regions of chicken BLBII and BLBIII genes, but greater than those for BLBII and BLBIII (not shown). Although interlocus divergence in some regions is high (figs. 4 and 6), overall, Agph-DAB1 and Agph-DAB2 and the DAB fragment have diverged less than have paralogous mammalian class II B genes, many of which are thought to have diverged before the diversification of extant placental mammals (Slade et al. 1994).

Ghost of Selection Past at Agph-DAB2

Despite the fact that Agph-DAB1 and Agph-DAB2 are closely related and the PBR region of Agph-DAB2 exhibits little polymorphism, the pattern of divergence from DAB1 in exons 2 and 3 reveals a history of divergent and stabilizing selection, respectively. In the 24 putative PBR codons (Brown et al. 1993), Agph-DAB1 and Agph-DAB2 have diverged ~4.2 times (Jukes-Cantor), or ~10.8 times (ML) more at nonsynonymous than at synonymous sites (fig. 6). Although interlocus divergence in some regions is high (figs. 4 and 6), overall, Agph-DAB1 and Agph-DAB2 and the DAB fragment have diverged less than have paralogous mammalian class II B genes, many of which are thought to have diverged before the diversification of extant placental mammals (Slade et al. 1994).

Discussion

We sequenced a 38.8-kb region in and around an Mhc class II B gene from a red-winged blackbird. Although much shorter than available sequences from the chicken and mammalian class II regions (Kaufman et al. 1999b), our blackbird sequence is the longest continuous DNA sequence reported thus far from perching birds (Passeriformes), a clade representing over half of all avian species, and offers a glimpse into the architecture of a songbird genome at the nucleotide level. The shotgun sequencing methods we used have not been widely employed in nonmodel vertebrate species and should be useful for understanding genomic organization, molecular evolution, and evolutionary relationships in birds and other vertebrates. We can therefore expect increasing application of such methods to nonmodel species in the future.

Comparison of Blackbird and Chicken Mhc Cosmid Sequences

Given that the chicken Mhc, the B complex, is extremely gene dense (Kaufman et al. 1999b), we were
surprised that our sequence contained but a single Mhc gene and little evidence for other functional genes. Some researchers prefer a functional definition of Mhc genes, such that a gene cannot be designated Mhc unless it is shown to be involved in graft rejection and linked to non-Mhc genes that are found in the Mhcps of model species (Kaufman et al. 1999a). We prefer a phylogenetic definition of Mhc genes, such that a gene is shown to be an Mhc gene (functional or nonfunctional) if it clusters with other Mhc genes to the exclusion of non-Mhc genes. Our Agph-DAB2 and the DAB fragment clearly satisfy this second criterion; thus, we are confident that we are in fact analyzing Mhc genes in blackbirds. Nonetheless, we do not yet know whether the cosmid we sequenced is linked to functional Mhc genes or to the majority of other Mhc sequences in the blackbird genome, of which there are many (Edwards, Nusser, and Gasper 2000). Thus, it is premature to suggest that the functionally significant portion of the blackbird Mhc is less compact than the functionally significant portion of the chicken B complex (Kaufman et al. 1999a, 1999b). Furthermore, the “minimal essential” chicken Mhc has been defined on both structural and functional grounds, with the latter being based on the observation of only one class I and one class II gene that are dominantly expressed (Kaufman 1995; Kaufman et al. 1999a). Nonetheless, this first glimpse at the blackbird class II region reveals an Mhc gene—containing region that is not as structurally streamlined as are currently sequenced Mhc-containing regions in the chicken (Kaufman et al. 1999b). A similar conclusion (with similar caveats) was reached upon examination of 32-kb in and around a house finch class II pseudogene (Hess et al. 2000). We do not yet know the primary structure of the other major Mhc regions in chickens, the Rfp-Y region. This region is known to contain expressed class I and II genes, as well as other non-Mhc genes (Miller et al. 1994). However, the total density and relative spacing of genes in this region are not yet known, making determination of the orthology of the blackbird sequence difficult.

Nonetheless, our sequence reveals some intriguing surprises that could aid in determination of orthology with model organism Mhcs after more characterization. Perhaps the most relevant to the search for orthology of the blackbird sequence is the presence of collagen-like and retinoic acid β-like sequences near Agph-DAB2. Kasahara et al. (1996) showed that the human class II region contains a collagen type V/XI gene (see Kasahara 1999 for a review). Although the collagen gene family in humans is large (~35 genes; Strachan and Read 1996), it is not so large that the some of the collagen fragments we have identified are real and potentiallyorthologous. More strikingly, a retinoic acid receptor β (RXRB) protein fragment was also implied in our cosmids. RXRB was also identified by Kasahara et al. (1996) in the human class II region as belonging to the group
of genes that make up an ancestral chromosomal duplication region containing multiple paralogs in the \textit{Mhc}. However, neither collagen nor retinoic acid receptor genes have been identified in the chicken \textit{Mhc} regions sequenced to date (Kaufman et al. 1999a, 1999b). Indeed, none of the non-\textit{Mhc} genes and gene segments we detected have homologs in the 92 kb of chicken \textit{Mhc} that has been sequenced to date, making validation of the orthology hypothesis difficult. These and other fragments identified in our analysis are intriguing but are thus far only fragments (table 1) and need to be followed up with more detailed phylogenetic analyses. It is likely that the cosmids we sequenced is a small part of a much more extensive class II region in this species. Consistent with this hypothesis is that the number of \textit{Mhc} class II hybridizing fragments in blackbirds as detected on Southern blots is large, making it likely that the number of actual class II B genes and gene fragments is larger than that in chickens (Edwards, Nusser, and Gasper 2000). Regardless of these possibilities, however, we have shown that at least one \textit{Mhc}-containing region in the blackbird genome is less streamlined and compact than the functionally important \textit{Mhc} region of chickens. In the event that the sequence of \textit{Rwcos10} proves to be representative of the functionally important \textit{Mhc} regions of blackbirds and other perching bird lineages, if the hypothesis that perching birds are a basal avian lineage may represent a derived rather than a primitive condition within birds. Some features of the sequence, such as the low density of long microsatellites, do conform more closely to rules emerging from chicken genomics (McQueen et al. 1996; Primmer et al. 1997). Avian genomes appear to be depauperate relative to mammals in SSRs, a feature consistent with their smaller genome sizes. At about one polymorphic (>15 repeats) microsatellite per 2 kb, a similarly sized region of the human class I or II region would have revealed at least 15 highly polymorphic microsatellites. We expect accumulation of further long sequences in songbirds to clarify any quantitative differences among avian species in SSR density and other features of genome architecture and to help determine how faithfully the chicken genome represents the characteristics of other avian genomes.

Evolution of \textit{Agph-DAB2}

Like some functional but poorly expressed chicken class II B genes, such as BLBIII, the level of diversity at \textit{Agph-DAB2} was low, much lower than diversity at the \textit{Agph-DAB1} PBR ($\pi = 0.101$, $\Theta = 0.070$) or intron 2 ($\pi = 0.018$, $\Theta = 0.040$) regions (Garrigan and Edwards 1999). The distribution among codon positions of the six segregating sites, two in each of the three codon positions, is consistent with \textit{Agph-DAB2} being a pseudogene. In conjunction with the negative and significant value of Tajima’s $D$ for \textit{Agph-DAB2}, it seems likely that diversity at \textit{Agph-DAB2} is not as strongly elevated by linkage to genes under balancing selection as regions close to \textit{Agph-DAB1} and some HLA-linked pseudogenes (Grimsley, Mather, and Ober 1998; Garrigan and Edwards 1999). The excess of rare variants suggested by Tajima’s $D$ could be explained if \textit{Agph-DAB2} was neutral but blackbird population size was increasing. Blackbird populations have been increasing dramatically in the United States over the last 50 years (less than 50 blackbird generations; G. Orians, personal communication), but this timescale is probably too short to affect the distribution of nucleotide diversity substantially. Rather, these data may indicate that \textit{Agph-DAB2} is relatively far away from genes under balancing selection or that the mutation rate of this gene is low, an uncertainty that could be clarified by examining the extent of divergence of \textit{Agph-DAB2} from genes in other species or other genes in blackbirds. In addition, we do not yet know the value of the neutral mutation parameter $\Theta$ for nuclear loci in blackbirds, a number that would clarify the dynamics of \textit{Agph-DAB2} considerably.

We can gain insight into evolutionary forces acting on blackbird genes by examining spatial variation in the amount of divergence between these loci, as in figures 4 and 6. Specifically, we can test the hypothesis that rates of silent change in different regions of the genes are the same, and thereby gain insight into mutational forces influencing the evolution of different exons and domains (Hudson, Kreitman, and Aguadé 1987). Population genetic theory predicts that levels of linked neutral divergence between two diverging paralogs, such as \textit{Agph-DAB1} and \textit{Agph-DAB2}, should be unaffected by balancing selection in the region (Hudson, Kreitman, and Aguadé 1987; Birky and Walsh 1988). This is because in a stationary population, the decrease in the rate of fixation of neutral linked sites due to balancing selection will be cancelled precisely by the increase in the number of neutral mutations per generation due to larger $N_e$ of balanced alleles. This logic depends on the genes being truly diverging; i.e., there is no evidence of “translocus” polymorphism (Imanishi 1995) such as would be expected to occur if alleles at the two genes had not yet achieved reciprocal monophyly, a pattern that we can reject for the PBR sequences (unpublished data). Thus, our finding of significant spatial heterogeneity in silent divergence between \textit{Agph-DAB1} and \textit{Agph-DAB2} (figs. 4 and 6) suggest that spatial variation in mutation or gene conversion rates may be important in the divergence of these genes. Interlocus gene conversion is thought to elevate silent rates in \textit{Mhc} PBR exons (Ohta 1998); such conversion may have contributed to the elevated level of divergence of \textit{Agph-DAB1} and \textit{Agph-DAB2} in both exon 2 and intron 1.

Birth-and-Death Process at Blackbird Class II Loci

The sequences in figure 5 are mostly the result of PCR amplifications of cDNA or genomic DNA that are not targeted to specific loci; therefore, it is conservative to consider many of the sequences gleaned thus far from songbirds different loci (Edwards et al. 1995). Thus it is premature to discuss issues of transspecies polymorphism of \textit{Mhc} alleles at specific loci in birds, because
we do not yet know which loci generate most of the sequences in the current database. Nonetheless, the species-specific clustering of sequences, also found in previous analyses of exon 3 (Edwards et al. 1995), is striking. We have previously interpreted such clustering as concerted evolution under the assumption that many of the sequences derive from different loci, rather than a lack of transspecies polymorphism (Edwards et al. 1995, 1999). Further genomic cloning of avian genes should provide a strong basis for analysis of allelic polymorphism at individual genes required to test the transspecies hypothesis.

Our sequence analysis provides a high-resolution view of a birth-and-death process of multigene family evolution in avian Mhc genes (Nei, Gu, and Sitnikova 1997; Gu and Nei 1999). In this model, there is constant turnover of genes by birth (duplication) and death (loss or pseudogene formation). The frameshift mutation in exon 2 appears to have arisen very recently, and the fact that it has arisen to an appreciable frequency suggests that even the alleles without a deletion are nonfunctional and neutral. The estimated origin of Agph-DAB2 and the blackbird class II B fragment at 40 MYA leads to the prediction that orthologs of these nonfunctional genes should be found in species of the songbird clade that diverged subsequent to this time, provided they have not been physically deleted. Songbird Mhc class II genes exhibit properties not only of the birth-and-death model of multigene family evolution, but also of the concerted-evolution model, in which frequent and extensive interlocus gene conversion or very recent gene duplications result in genes clustering by species in phylogenetic trees (Edwards et al. 1999). The Mhc class II B sequences we have characterized here also support the two-model scenario (Nei, Gu, and Sitnikova 1997). The Mhc class II B pseudogenes described here and in the house finch (Hess et al. 2000) are among the only avian Mhc pseudogenes characterized to date and further support the claim that the class II regions of songbirds are less streamlined than those of chickens.

Some models for the evolution of multigene families predict a period of relaxed and, in some cases, divergent selection on novel genes just after duplication as they either degenerate into pseudogenes or acquire new functions (Ohta 1991; Hughes 1994). The pattern of divergence between Agph-DAB1 and Agph-DAB2 sequences, as well as inferences of paths of evolution at PBR sites from ancestral sequences, suggests an episode of divergent selection acting on PBR sites of both genes.

The fact that a similar pattern of interlocus divergence is found in comparisons of the blackbird and house finch pseudogenes indicates that the pattern seen in the Agph-DAB1/Agph-DAB2 comparison is not a result of divergent selection acting solely on the highly polymorphic Agph-DAB1. Thus, despite its current status as a pseudogene, the PBR of Agph-DAB2 apparently diverged adaptively away from genes such as Agph-DAB1 sometime after duplication (fig. 6).

We attempted to use $d_{S}/d_{D}$ ratios to estimate the time when Agph-DAB2 became nonfunctional (Miyata and Yasunaga 1981), but this method requires stringent assumptions that our data did not fulfill. Nonetheless, the signal implicating a past period of adaptive evolution in the blackbird genes—a ghost of selection past—is particularly strong. A similar pattern of divergence has been documented at functional mammalian class II B genes, but in most cases these comparisons involve genes that diverged prior to the diversification of eutherian lineages, and the resulting indices of adaptive divergence ($d_{S}/d_{D}$ ratios) are often low, suggesting saturation at PBR sites (Hughes and Nei 1989). A “ghost of selection past,” often used to describe organismal evolution but implicit in some models of multigene family evolution, is invoked to describe situations in which the footprint of selection by an extinct organismal agent, such as a pollinator or seed disperser, is still evident in extant species with which it interacted (Janzen and Martin 1982). This ghost is all the more evident at blackbird Mhc genes because of the recency of origin of Agph-DAB2. Apparently, there is a fairly high frequency of trial and error in the duplication process of blackbird Mhc class II B genes, a scenario that could characterize other songbird species and Mhc regions.

Acknowledgments

We thank D. Westneat for blackbird DNAs, M. Lee, B. Ewing, and N. Takezaki for computational assistance, J. Kaufman, H. Wichman, and Y. Satta for helpful discussion, and T. Ohta, C. Hess, two anonymous reviewers, and C. Aquadro for comments on the manuscript. We thank B. Paine, G. Orians, and D. Futuyma for clarifying the distinction between “ghost of competition past,” oft used in community ecology, and “ghost of selection past,” a conceptually old but idiomatically new derivative. This work was supported by NSF grants DEB9707548 and 9815800 to S.V.E. and NSERC grants to B.F.K.

LITERATURE CITED


Charles Aquadro, reviewing editor

Accepted May 31, 2000