 doi:10.1006/geno.2001.6596, available online at http://www.idealibrary.com on IDEAL

Songbird Genomics: Analysis of 45 kb Upstream of a Polymorphic *Mhc* Class II Gene in Red-Winged Blackbirds (*Agelaius phoeniceus*)

Joe S. Gasper,[1] Takashi Shiina,[2] Hidetoshi Inoko,[2] and Scott V. Edwards[1,*]

[1]*Department of Zoology, Box 351800, University of Washington, Seattle, Washington, 98195, USA*
[2]*Department of Genetic Information, Division of Molecular Life Science, Tokai University School of Medicine, Bohseidai, Isehara, Kanagawa 259-11, Japan*

[*]*To whom correspondence and reprint requests should be addressed. Fax: (206) 543-3041. E-mail: sedwards@u.washington.edu.*

Here we present the sequence of a 45 kb cosmid containing a previously characterized polymorphic *Mhc* class II B gene (*Agph-DAB1*) from the red-winged blackbird (*Agelaius phoeniceus*). We compared it with a previously sequenced cosmid from this species, revealing two regions of 7.5 kb and 13.0 kb that averaged greater than 97% similarity to each another, indicating a very recent shared duplication. We found 12 retroelements, including two chicken repeat 1 (CR1) elements, constituting 6.4% of the sequence and indicating a lower frequency of retroelements than that found in mammalian genomic DNA. *Agph-DAB3*, a new class II B gene discovered in the cosmid, showed a low rate of polymorphism and may be functional. In addition, we found a *Mhc* class II B gene fragment and three genes likely to be functional (encoding activin receptor type II, a zinc finger, and a putative γ-filamin). Phylogenetic analysis of exon 2 alleles of all three known blackbird *Mhc* genes indicated strong clustering of alleles by locus, implying that large amounts of interlocus gene conversion have not occurred since these genes have been diverging. Despite this, interspecific comparisons indicate that all three blackbird *Mhc* genes diverged from one another less than 35 million years ago and are subject to concerted evolution in the long term. Comparison of blackbird and chicken *Mhc* promoter regions revealed songbird promoter elements for the first time. The high gene density of this cosmid confirms similar findings for the chicken *Mhc*, but the segment duplications and diversity of retroelements resembles mammalian sequences.

Key Words: duplicon, perching bird, intron, concerted evolution, gene conversion

INTRODUCTION

The major histocompatibility complex (Mhc) is encoded by a large multigene family involved in the humoral and T-cell-mediated immune responses of vertebrates. Class I and II genes encode glycoproteins that transport foreign peptides to the surface of cells for recognition by T-cell receptors on lymphocytes, which in turn kill infected cells [1]. The peptide binding region (PBR) of class II Mhc molecules is encoded by two genes and binds exogenous antigens, whereas the class I PBR is encoded by a single gene and binds endogenously replicating antigens such as viruses. The PBR of *Mhc* genes is the most polymorphic region of any vertebrate gene family; this polymorphism is manifested as high heterozygosity and large divergences between alleles and is thought to increase the fitness of the organism against pathogen variation [2]. This molecular dynamism required for disease resistance by means of *Mhc* polymorphism may also explain the redundancies, frequent gene duplication, and pseudogene formation in *Mhc*; for example, the human leukocyte antigen (HLA) class II region contains approximately 40 genes, of which several are pseudogenes and gene fragments [3].

Segment duplications and retroelements such as long interspersed nuclear elements (LINEs) and human endogenous retroviruses (HERVs) have been observed frequently in both coding and noncoding regions of the HLA. In the class I region of humans, for example, Peri-B and Peri-C are 53-kb and 48-kb duplicons, respectively, containing HLA-B and -C coding regions [4]. Often such large duplicons contain paralogous genes, pseudogenes, gene fragments, and retroelements that permit estimation of the time of the

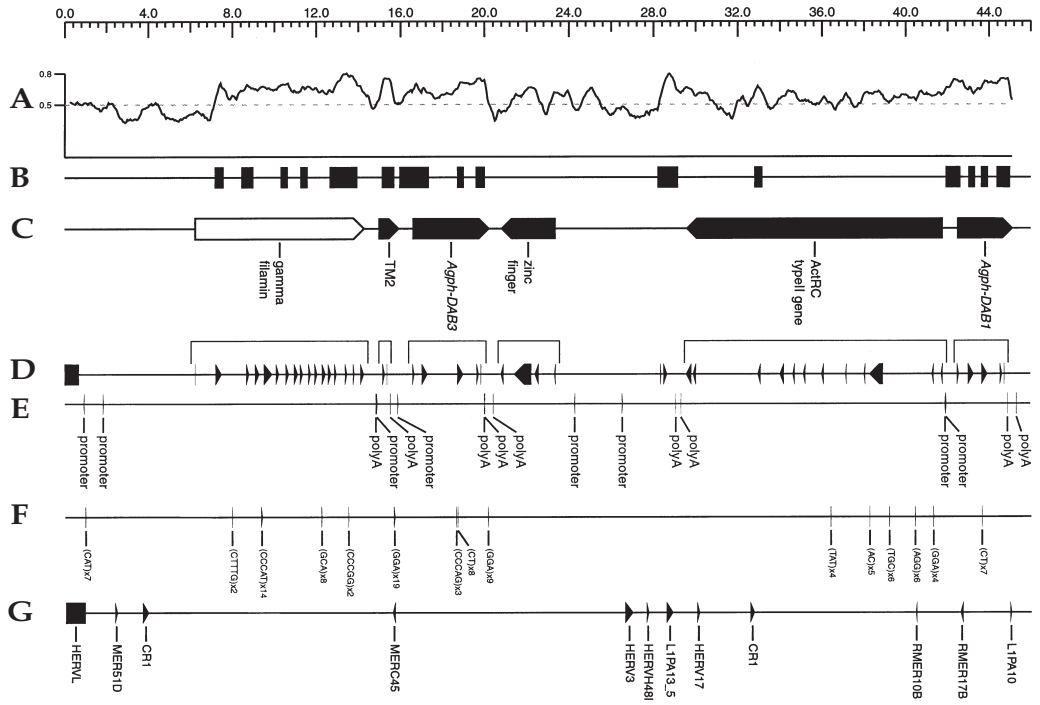

FIG. 1. Organization of red-winged blackbird cosmid 3. (A) Sliding window of %GC over entire cosmid using 500-bp windows and 50-bp offset length. (B) CpG islands predicted by SeqHelp. (C) Genes determined by internal programs of SeqHelp. The filled boxes are considered functional (excluding the *Mhc* fragment) and the open box indicates tentative assignment. (D) Exons predicted by Genscan. Brackets indicate exons of genes in (C). (E) Poly(A) tails and promoter regions found in the cosmid by Genscan. (F) Simple sequence repeats. (G) Retroelements predicted by Censor and RepeatMasker. The length of the genes (in base pairs) are as follows: γ-filamin, 8039; *Agph-DAB3*, 3666; zinc finger, 2598; activin receptor type II, 12214; *Agph-DAB3*, 2681.

duplication events and their subsequent diversification. Retroelements constitute more than 47% of the *Mhc* class I region in humans [5], a number that is higher than the predicted value of 35% for the entire genome [6]. Repetitive interspersed retroelements may facilitate evolutionary change through deletions or duplications within multigene families by means of homologous but unequal crossing-over, reverse-transcription-mediated translocations, or non-homologous chromosomal rearrangements. Thus, retroelement-mediated genomic instability may be a major mechanism by which redundancy in mammalian *Mhc* was acquired.

In contrast to results for mammals, sequencing of the chicken *Mhc* (B complex) has uncovered a highly streamlined multigene family. The 44 kb region in the center of the B complex is extremely compact, containing 11 genes (intergenic distances as small as 30 bp), but no pseudogenes or gene fragments [7]. These genes have smaller introns than their mammalian counterparts, and most of the genes have homologues in the mammalian *Mhc*, suggesting a "minimal *Mhc*" [8]. So far, however, there has been no published characterization of repetitive elements, segment duplications, or tandem repeats in chicken *Mhc* other than a repeat of a single lectin gene outside the central class II region. The entire chicken B complex is contained within 92 kb with little spatial segregation of class I, II, and III regions. This paucity of repeats and the compactness of the region is consistent with experimental evidence for little or no recombination [9], and the entire B complex is thought to co-evolve as haplotypes, with alleles

	S		X		Y	matches with consenseus		
	GGACCTY	Pyr Tract	CCYAGMRACAGATG	X2 Sp-1	CTGATTGGYY	S	X	Y
BLBII	...G.C.	CGCGGCGCAGAACTCTG	...G.....G.G..	ACGCCGCCCGGCGCCGCCGCG	..C.....	5	11	9
Agph-DAB1	..G..AG	..TCA.CG----...GC	.A...GC..C.C..	.GAT.C.G..C.C.T.GGCA--	.G.....C.G	4	9	7
Agph-DAB2	..G..AG	..TCA.CG----...GC	.A...GC..C.C..	.GAT.C.G..C.C.T.GGCA--	.G.....C.G	4	9	7
Agph-DAB3	A.C.AA.	G.AA.AC.-.GGG.G.A	T.GC...T..C...	TT.-.CAGGTAGAGG.TTCA--	GAA....AG.	3	9	5
Mouse EB	......G	AAACT.C.--T....G.	A.....T....	.T..T.GA.TC.TTT.ATG---	....	6	12	10
Human DQ_α	A.G....	TTAATACA.ACT..TCA	G.....GA	TGT.AC.AT..G.GATTTTT--	..A.....	5	9	9
Human DP_α	.C.....	.CA.C.T--CCT..T.A	.....GA	.T.T.AG.TCTATGATTTCT--	.....A...G	6	12	8

FIG. 2. Comparison of songbird *Mhc* class II B gene promoter regions with *B-LBII* (chicken class II), Mus E B (mouse class II), and HLA-DQ A and DP A (human class I) *Mhc* promoters. Within promoter motifs S, X, and Y (boxed), dots indicate matches with the consensus sequence (at top); between motifs dots indicate matches with the chicken sequence. Underlines upstream of X box are pyrimidine tracts and within or overlapping the right end of the X box are X2 motifs (protein binding factors of the following sequence: TGAG, TGAC, GTAC, or TGTCA). Putative Sp-1 binding sites (between but not overlapping the X and Y boxes) are underlined. Sequences are from the following sources: *B-LBII* [18], Mus E B [48], HLA-DQ A [49], and HLA-DP A [50]. Degenerate nucleotide symbols are as follows: Y, C or T; M, A or C; and R, A or G.

doi:10.1006/geno.2001.6596, available online at http://www.idealibrary.com on IDEAL

TABLE 1 Summary of retroelements recovered from blackbird cosmid 3

Retrotransposon element type	Retrotransposon database and organism match	Location in cosmid[a]	Length[a]	% Similarity[a]	Alignable region
retrovirus	HERVL / human	0-946	946	35.0 (protein level)[b]	partial gag gene
retrovirus	MER51D / human	2370-2495	115	79.0	partial 3′ LTR
LINE	CR1/chicken	3691-3986	295	73.0	partial RT ORF
DNA transposon	MERC45 / human	15586-15712	126	73.0	partial 5′ ITR
retrovirus	HERV3 / human	26614-26995	381	65.2	partial envelope gene
retrovirus	HERVH48I / human	27631-27745	114	65.7	partial envelope gene
LINE	L1PA13_5 / human	28583-28902	319	67.9	partial 5′ sequence
retrovirus	HERV17 / human	30044-30174	130	77.7	partial envelope gene
LINE	CR1/chicken	32557-32765	208	64.0	partial RT ORF
retrovirus	RMER10B / rodent	40450-40513	63	81.6	partial 5′ LTR
retrovirus	RMER17B / rodent	42580-42724	144	69.3	partial 3′ LTR
LINE	L1PA10 / human	44977-45058	81	73.4	partial ORF2

[a]Landmarks and % similarity in nucleotides.
[b]Detected by amino acid similarity using SeqHelp.

at different loci maintaining strong disequilibrium over long periods of time. The *Mhc* region corresponding to the chicken B complex of another galliform bird, the Japanese quail (*Coturnix coturnix*), was recently characterized and shown to be larger (150 kb) and less streamlined than that of the chicken, containing several duplications of the B-G, B-lectin, and NKR genes (T. S. *et al.*, manuscript in preparation).

The only other avian clade whose *Mhc* genes have been studied at the genomic level are the songbirds (oscines), part of a larger clade of perching birds (Passeriformes) containing about 5,300 species or over half of all extant birds [10]. We have focused our attention on the *Mhc* of the songbird clades because they are phylogenetically distant from gamebirds and because they exhibit behaviors that may be mediated by *Mhc* variation. Previously we have shown that genomic regions containing songbird *Mhc* genes possessed a small number of relatively short simple-sequence repeats (SSRs) and higher and lower frequencies of pseudogenes and functional genes, respectively, compared with chicken, making such regions structurally intermediate between those of mammals and chicken [11,12]. We have yet to characterize a genomic region surrounding a highly polymorphic and presumably functional avian *Mhc* gene. *Agph-DAB1* represents one such blackbird gene whose structure and polymorphism has been characterized [13,14]. Here we present the sequence of the cosmid bearing *Agph-DAB1* (Rwcos3) and describe two segment duplications shared with another sequenced cosmid, Rwcos10 [11], a new blackbird *Mhc* class II gene, *Agph-DAB3*, and other coding and noncoding regions that provide further details of the evolution of blackbird *Mhc* genes and their genomic neighborhood.

RESULTS

Features of the Blackbird Sequence

The sequence of Rwcos3 was 45,375 bases, making it the longest sequence obtained thus far from a non-game bird (Fig. 1). The percentage of GC (%GC) of the cosmid varied from 33.0% to 78.6% in 500-bp windows (Fig. 1A), averaging 54.5% over the entire length. We found 15 CpG islands, which are often good indicators of genes in chicken genomes [15], in three clusters (Fig. 1B). In addition to *Agph-DAB1*, a second full-length *Mhc* gene (*Agph-DAB3*) was identified using SeqHelp, along with a *Mhc* gene fragment (Fig. 1C) designated transmembrane fragment 2 (TM2) because it is the second such fragment that aligns with the transmembrane-encoding exons 4–6 of *Mhc* class II B genes. *Agph-DAB1*, *Agph-DAB3*, and TM2 were in regions with greater than 65% GC content. SeqHelp also identified two genes, encoding a zinc finger domain and an activin receptor type II (Fig. 1C), both with inferred coding sequences uninterrupted by in-frame stop codons. The zinc finger domain, of the *Krüppel* Cys2His2-type thought to regulate development through the DNA binding properties of the C2H2 finger repeats [16], was composed of four exons (Fig. 1D) and possessed a promoter region identified by Genscan as well as a poly(A) site (Fig. 1E). Exon 3, the longest of the four exons, contained eight C2H2 repeats. The closest GenBank match, a *Krüppel*-like zinc finger on human chromosome 9, also has 4 exons with 16 C2H2 repeats in the third exon. The activin receptor type II gene, which encodes transmembrane glycoproteins, was determined to have 12 exons with a promoter region and a poly(A) site (Fig. 1E). The closest GenBank match was a mouse activin receptor type II gene, which has 11 exons

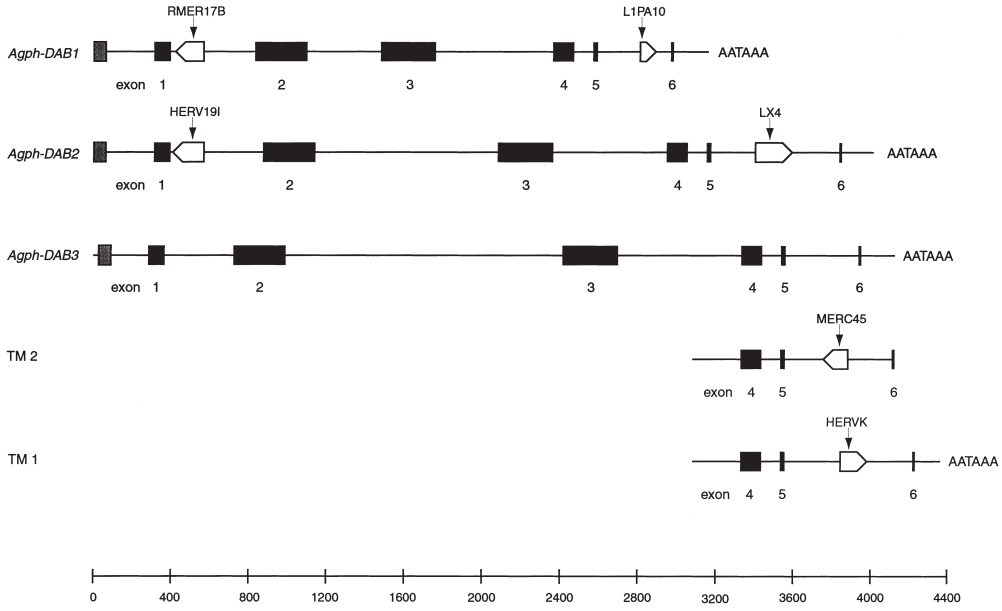

FIG. 3. Size and structure of blackbird *Mhc* class II B genes and *Mhc* fragments, including promoter regions (gray shaded boxes) and poly(A) sites. Six retroelements are indicated as open boxes in introns 1 and 5 of *Agph-DAB1* and *Agph-DAB2*, and in intron 5 of *Mhc* fragments TM2 and TM1.

with 2 very long introns making up most of the length (> 66 kb) [17]. Similarly, 2 introns totaling 5480 bp of the predicted blackbird activin receptor gene make up more than 45% of its length. Also identified using SeqHelp were 11 short (< 30 bases) DNA segments matching a human γ-filamin gene. The DNA matches for the γ-filamin gene total 167 bases with a 97.9% GenBank match over a 4.9-kb span, suggesting a possible gene (Fig. 1C). SeqHelp predicted five CpG islands in that region (Fig. 1B) in a single cluster, again raising the possibility of a gene. Additionally, Genscan identified a full gene with 17 exons in the same region (Fig. 1D), but we were unable to find any matches of the predicted protein in the GenBank database using Blast.

Overall, Genscan identified 49 exons making up the 4 genes discussed above, *Agph-DAB1*, and TM2 (Fig. 1D, bracketed). Eight of these predicted exons corresponded exactly to manual predictions of exons 1–4 of both *Agph-DAB1* and *Agph-DAB3*. Exons 5 and 6 may have been missed in both genes due to their short lengths (34 and 14 bp, respectively; Chris Burge, Cambridge, Massachusetts, pers. comm.). Also, exon 4 of TM2 was correctly predicted. The exon predicted on the 5′ end of the cosmid (Fig. 1D) aligns with a human endogenous retrovirus *gag* gene (encoding the viral structural capsid proteins; Table 1) and the two exons upstream of the activin gene produced no BLAST hits at either the DNA or the protein levels. A total of 15 microsatellites, varying in length from 10 to 70 bp and 2 to 19 repeats, were recovered using the program Sputnik (Chris Abajian, unpublished data; Fig. 1F).

Mhc Gene Promoter Regions

Genscan did not specifically identify a typical classical *Mhc* class II gene promoter region of either *Mhc* gene in the sequence, so we searched for them by eye, using the chicken promoter regions as guides [18]. The promoter regions identified for *Agph-DAB1*, *Agph-DAB2*, and *Agph-DAB3* (Fig. 2) were found 375 bp upstream of the start codon for *Agph-DAB1* and *Agph-DAB2*, and 346 bp for *Agph-DAB3*. They have a structure similar to those of the chicken *Mhc* class II genes, with clear S, X, and Y boxes (transcription controlling sequence motifs) [19]. *Agph-DAB1* and *Agph-DAB2* have 20 of the 31 (64%) bases in the S, Y, and X box consensus sequence, whereas *Agph-DAB3* has 17 (57%; Fig. 2). All three DAB genes share conserved lengths between the X and Y boxes of 19–22 bp and have X2 motifs (known to be important for promoter function) of 4–8 bp immediately downstream of the X box. A motif similar to X2 of *Agph-DAB3* was found inside the X box. *Agph-DAB1* and *Agph-DAB2* both possess pyrimidine tracts known to be conserved in *Mhc* class II promoter regions of chickens and mammals and have a Sp-1 binding site with 1-bp difference (CCGCGC) from that of chickens (CCGCCC; Fig. 2). All of the DAB genes have the inverted CAAT motif in the Y box found in chicken promoters and the invariant G at the fifth position in the X box, which is conserved throughout chicken and mammalian *Mhc* class II promoter regions.

Retroelements

The program Censor identified nine retroelements; RepeatMasker recovered an additional three. Four of the retroelements were LINEs, all of which were fragmentary, ranging in size from 81 to 319 bp. Two of the LINEs aligned with CR1 elements found in chickens [20], consisting of partial reverse transcriptase open reading frames (ORFs; Table 1). In addition, we found seven fragmentary retroviruses. A sequence similar to a HERVL-type retrovirus identified at the extreme 5′ end of the sequence was the longest retroelement found in the cosmid, extending 946 bp and encoding a near full-length *gag* gene (Fig. 1G). Alignable but truncated

doi:10.1006/geno.2001.6596, available online at http://www.idealibrary.com on IDEAL

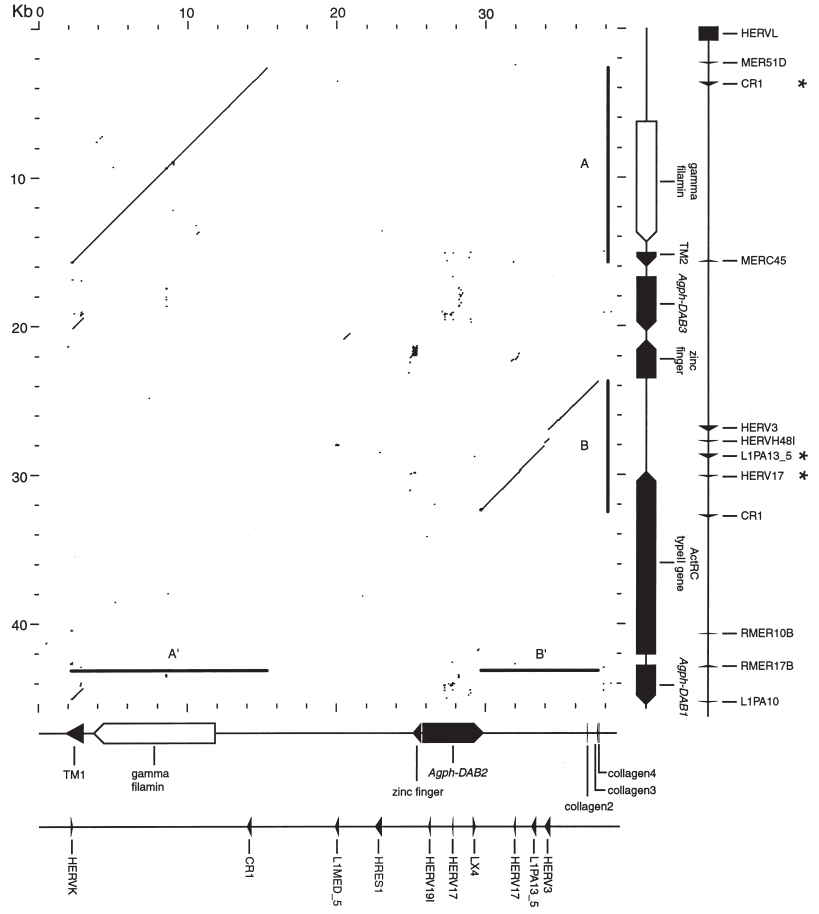

FIG. 4. Dot-plot analysis of cosmid10 (x axis) compared with reverse complemented cosmid 3 (y axis) in which each dot represents a match of 33 nucleotides per 50. The long diagonals show segment duplicons between the cosmids and retroelements marked with an asterisk are shared between the duplicated segments. Homology between *Agph-DAB1*, *Agph-DAB3*, and *Agph-DAB2* is not indicated due to their opposite orientation on the cosmid sequences. Note putative γ-filamin gene on cosmid 10 not previously reported [11].

retrovirus elements of the mammalian types RMER10B and HERV19I, 144 and 160 bp, respectively, were identified 12 nucleotides into intron 1 of *Agph-DAB1* and 3 nucleotides into intron 1 of *Agph-DAB2* (Fig. 3). Similarly, partial LINEs similar to human L1PA10 (81 bp) and LX4 (180 bp) elements were found 208 bp into intron 5 of *Agph-DAB1* and *Agph-DAB2*, respectively. TM2 contains an element in intron 5 similar to the human MER45C DNA transposon, which inserts into the genome without an RNA intermediate, and TM1, a *Mhc* class II fragment on Rwcos10, has a retrovirus (HERVK type) also inserted in intron 5. The 12 retroelements totaled 2,922 bases in length and composed 6.4% of the cosmid sequence with an average of 71.8% nucleotide similarity to database matches.

Divergence of Duplicated Segments

A dot-plot analysis between Rwcos10 and Rwcos3 identified two duplicons on each, resulting in four segments, designated A (Rwcos3), A′ (Rwcos10), B (Rwcos3), and B′ (Rwcos10; Fig. 4). The A and A′ segments were both exactly 13,035 bp, with no apparent compensatory insertions or deletions, and differed at 231 sites, resulting in 98.4% similarity. The B and B′ segments were 8,409 and 7,563 bp, respectively, with the retroelement type HERVH48I making up 114 bp of the 846 bp indel in segment B (Fig. 4). Excluding the indel, the B and B′ segments were 96.8% similar with 229 differences. Segments A and A′ each contained a copy of a region identified as the γ-filamin gene and also a copy of a *Mhc* gene fragment described previously. Using a less diverse set of search engines, a series of unrelated gene fragments were identified in the A′ region [11], including a retinoic acid B gene, but we now favor the hypothesis that this region contains a γ-filamin gene. We examined the segments for shared retroelements in paralogous positions that might suggest pre-duplication insertion events [4] and found three, the HERV17 type and the L1PA13_5 type shared by segments B and B′, and the CR1 element shared by segments A and A′ (Fig. 4, asterisk). We estimated the divergence between A and A′ (d = 0.016 ± 0.002) to be 8 million years ago (MYA) ± 1.0 million years (MY) and the divergence between B and B′ (d = 0.032 ± 0.002) to be 16.2 MYA ± 1.0 MY.

Structure, Polymorphism, and Phylogenetics of *Agph-DAB3*

Agph-DAB3 is 3,666 bases from start to stop codon and has a poly(A) site 142 bases downstream of the stop codon (Fig. 3). Intron 2 of *Agph-DAB3* (1423 bp), typically the longest intron of songbird *Mhc* class II genes, is 3.7 and 1.5 times longer than intron 2 of *Agph-DAB1* and *Agph-DAB2,* respectively (Fig. 3), making *Agph-DAB3* the longest avian *Mhc* sequence found so far. A full-length inferred coding sequence of 264 amino acids showed 22 differences in exon 2 from *Agph-DAB1* and 29 from *Agph-DAB2*, and possessed all 19 conserved amino acid residues considered important for class II gene function (data not shown) [21], consistent with *Agph-DAB3* being a functional gene. To further investigate the functional status of *Agph-DAB3*, we surveyed polymorphism in exon 2 (encoding the PBR). Highly expressed *Mhc* class II B gene PBRs tend to have high nucleotide polymorphism, higher nonsynonymous to synonymous (d_N/d_S) rates of substitution, and higher numbers of alleles segregating in populations. *Agph-DAB3* exon 2 exhibited low polymorphism with only four segregating

sites among the six alleles found and an overall nucleotide diversity (π) of 0.0026. Tajima's *D* statistic was –0.852 and not significantly different from the value for a neutral locus with the same number of segregating sites ($P > 0.10$). The d_N/d_S ratio was 0.50, much lower than the ratio estimated for *Agph-DAB1* (~3) [14] and the chicken *B-LBII* locus (1.13) [22], both under strong balancing selection, but higher than the ratio for the chicken *B-LBIII* locus (0.28), a non-polymorphic but functional and expressed *Mhc* gene [23].

Exon 3 is highly conserved in *Mhc* class II genes and is not under balancing selection, thus providing a better estimate of divergence times than does exon 2. The divergence of *Agph-DAB1* and *Agph-DAB3* ($d_S = 0.062 \pm 0.031$) is estimated to have occurred approximately 31.0 MYA ± 15.5 MY. The divergence of *Agph-DAB2* and *Agph-DAB3* ($d_S = 0.063 \pm 0.032$) occurred approximately 31.5 MYA ± 16.0 MY, and the divergence of *Agph-DAB1* and *Agph-DAB2* ($d_S = 0.031 \pm 0.022$) about 15.5 MYA ± 11.0 MY. The high standard errors of the estimates are likely due to the low divergence and the low number of silent sites in the comparisons. A neighbor-joining tree of alleles of all three of the blackbird loci indicated that alleles of each locus strongly clustered together (all bootstraps > 90%; Fig. 5); however, the sequences provided only weak support for a sister-gene relationship of *Agph-DAB1* and *Agph-DAB3*. The higher polymorphism of *Agph-DAB1* alleles relative to *Agph-DAB2* and *Agph-DAB3* is clearly evident from the relative coalescence times of alleles for each locus in the tree (Fig. 5).

FIG. 5. Phylogenetic analysis of the exon 2 alleles from blackbird genes *Agph-DAB1*, *Agph-DAB2*, and *Agph-DAB3*. The tree was built by the neighbor-joining method using Kimura 2 parameter distances [45]. Numbers indicated are bootstrap percentages from 2000 replicates.

DISCUSSION

We have characterized a 45-kb sequence from a red-winged blackbird and analyzed in detail the presence and distribution of coding regions, retroelements, and block duplications. Although not very long by mammalian genomics standards, the blackbird sequence is one of a very few of its kind for birds. In addition to recovering a novel *Mhc* class II B gene, our analysis has revealed an impressive diversity of retroelements and duplicons, more so than have been described for existing chicken and quail sequences [7,24]. These duplications are clearly not allelic to previously characterized blackbird sequences, and so are best interpreted as paralogous rather than homologous to other regions of the blackbird genome.

Function, phylogenetics, and origin of *Agph-DAB3*

Agph-DAB3 is a full-length *Mhc* class II B gene possessing the expected six exons, five introns, an identifiable promoter region, and a poly(A) tail. The location, sequence, and spacing of the promoter region of *Agph-DAB3* does bear similarity to known functional X and Y boxes deemed the most important structures for promoter function in knockout experiments [19]. There was, however, no pyrimidine tract upstream of the X box or a discernable Sp-1 binding site found between the X and Y boxes as found in *Agph-DAB1*, *Agph-DAB2*, and the highly polymorphic *B-LBII* locus in chickens. Also, the promoter region of *Agph-DAB3* differed from chicken and mammalian *Mhc* promoters with only three matches with the consensus S. This homology of the S box seems tenuous, but 5′ deletion analysis revealed that this region is less important for promoter function [19]. The promoter region of the *B-LBIII* locus of the B12 haplotype from a chicken lacks a homologue of the S box and has two mutations in the core of the Y box [18], but continues to be expressed at low levels [23]. Perhaps *B-LBIII* and *Agph-DAB3* are in the process of degeneration; alternatively, their non-polymorphic PBRs may target highly conserved epitopes of certain pathogens, as in some non-classical mammalian *Mhc* genes [25]. Consistent with this second hypothesis, the d_N/d_S ratio of the PBR is less than 1, implying weak stabilizing selection. This last suggestion indicates that blackbird *Mhc* genes experience a diversity of types of selection, from balancing (*Agph-DAB1*) to stabilizing (*Agph-DAB3*) to loss of function (*Agph-DAB2*).

Avian Mhc genes tend to cluster phylogenetically by species rather than by gene, suggesting either recent gene duplications or higher rates of interlocus gene conversion than that of mammals [26]. Agph-DAB1, Agph-DAB2, and Agph-DAB3 haplotypes formed monophyletic clades, implying that if interlocus gene conversion does occur, it does so at a level low enough or far enough in the past to

doi:10.1006/geno.2001.6596, available online at http://www.idealibrary.com on IDEAL

TABLE 2 Comparison of three Mhc gene bearing cosmid sequences from songbirds with that of the B locus from chickens

Cosmid and organism	%GC	kb	Number of genes (per 10 kb)	SSR (>15bp) (per 10 kb)	Retroelements (per 10 kb)	% DNA unaccounted (excluding introns)	Reference
cosmid10 / blackbird	54.5	38.7	2 (0.52)	3 (0.77)	10 (2.5)	58.8	Edwards *et al*., 2000
cosmid 3 / blackbird	55.0	45.3	5 (1.1)	9 (2.0)	12 (2.6)	26.9	this paper
total blackbird cosmids	54.7	84.0	7 (0.83)	12 (1.4)	23 (2.7)	42.8	— — —
cosmid10A / house finch	56.9	31.9	2 (0.62)	7 (2.1)	7 (2.2)	86.0	Hess *et al*., 2000
class II region of BF/BL locus/chicken[a]	62.0	26.9	7 (2.3)	5 (1.7)[b]	0	37.0	Kaufman *et al*., 1999a

[a]The class II region includes the segment from the BLb1 gene to the BMb1 gene, bp 33008 to 59992 [7], and was analyzed for this paper.
[b]SSR data was calculated with the program Sputnik.

preserve phylogenetic integrity of extant alleles at these loci. However, the blackbird *Mhc* genes are still very closely related to one another compared with typical paralogous mammalian class II genes, even those in the same subfamilies, such as human *DRB1* and *DRB2*; many mammalian paralogous *Mhc* genes are thought to have diverged over 60 MYA. The bird results are consistent with a model in which duplications or gene conversion events occurred after the divergence of blackbirds and other songbird species sampled so far [11,12].

The tree in Fig. 5 indicates that *Agph-DAB2* is the descendant of an initial duplication event before the duplication that gave rise to *Agph-DAB1* and *Agph-DAB3*. By contrast, the occurrence of possible homologous retroelements in *Agph-DAB1* and *Agph-DAB2* suggest that these genes may be more closely related to one another than to *Agph-DAB3*. Homologous retroelements have often been used as phylogenetic markers to infer the order of duplication events [27], as integration into the same location is unlikely to occur convergently and is irreversible. The low bootstrap value placing *Agph-DAB1* and *Agph-DAB3* as a sister group (60%) indicates that the sequences exhibit a muted phylogenetic signal, perhaps due to homoplasy, ancient interlocus gene conversion, or an inappropriate outgroup. The retroelements shared between duplicons B and B′ (HERV17 and L1PA13_5 element types) and duplicons A and A′ (CR1) are further examples of retroelements as indicators of common ancestry.

Mammal-like Retroelement-Mediated Duplications?

In the class I-α block region of the human *Mhc* there are 10 tandem segment duplicons, each up to 50 kb, with interspersed repetitive elements making up to 29% of sequence. The shorter duplicons observed in our sequences (< 13.5 kb) may be related to the smaller genome size of birds relative to mammals or other processes that minimize the amount of junk DNA and redundancy in avian genomes [28]. Consistent with this are the low frequency of microsatellites, which are absent from the chicken B complex and are estimated to compose only 0.44% of the chicken genome [29], and are also infrequent in the combined blackbird cosmids (0.78%). In addition, 22 retroelements were identified across cosmids Rwcos3 and Rwcos10. This density of retroelements (6.35%) is much lower than mammals, and is also lower than estimated for chickens (17%) [29]. Based on their occurrence near the ends of duplicons, repeat elements have been proposed to help duplicate *Mhc* segments as well as generate multiple *Mhc* gene fragments [4,30]. Although we did not find evidence for these mechanisms in the blackbird cosmids, it is possible that duplicons larger than those currently available were undetected in this study but are facilitated by repeats.

Evidence suggests a more recent ancestry for duplicons A and A′ than for B and B′, in addition to the twofold more recent estimate of divergence time for A and A′ from sequence comparisons. As expected for older duplications, segments B and B′ have accumulated more post-duplication retroelement insertions, compared with no insertions in segments A and A′. The inferred amino acid sequences of TM2 and TM1 and the γ-filamin genes shared between the A and A′ duplicons are nearly identical (data not shown). These results imply a history of duplication events occurring at different times. In conjunction with the divergence times of the blackbird *Mhc* genes, the segments imply three tiers of duplication, approximately 8 MYA (segment A and A′), 16 MYA (segment B and B′, *Agph-DAB1*, and *Agph-DAB2*), and 31 MYA (*Agph-DAB3* and *Agph-DAB1*). These divergence times describe divergence since duplication or since the last round of interlocus gene conversion. Although exon 3 is subjected to interlocus gene conversion and concerted evolution [31], which will bias the estimated dates of divergence [32], we expect this to be minimal because the interlocus phylogeny bears no evidence of frequent gene conversion in the recent ancestry of these genes.

Comparison with chicken B locus

The two class II-bearing blackbird cosmids we have characterized thus far, totaling 84 kb, are nearly as long as the entire *Mhc* B complex in chickens (92 kb). None of the non-*Mhc* genes recovered from our blackbird sequences (zinc finger, activin

receptor type II, and the γ-filamin) occur in the B complex and no retroelements have been detected, compared with 22 in our songbird sequences. These patterns indicate that orthology does not exist or that the evolution and structure of the *Mhc* regions between chickens and songbirds is vastly different. Alternatively, the higher density of retroelements in the blackbird sequences may indicate a genomic region undergoing degeneration away from the typical pattern of genomic streamlining observed in chickens. Table 2 does imply some similarities between the Rwcos3 sequence and the chicken class II region of the B complex; for example, the SSR density, the average GC content, and the gene density are roughly the same. Clearly, however, the limited sequencing in *Mhc*-containing regions of songbirds already reveals substantial structural diversity and divergence from the B complex. Similarly, *Mhc* sequences from a close relative to chickens, the Japanese quail, is nearly twice as long (150 kb) and less streamlined than the B complex, with six duplicated B-G genes, three NKR genes, and six B-lectin genes. It remains to be seen whether the B complex is representative of other avian *Mhc* genes. Like the mammalian *Mhc*, the picture emerging from these first fragments of avian *Mhc* genomics is one of a very dynamic region with frequent duplication of genes and gene fragments. Identification and sequencing of songbird contigs, particularly those containing single-copy genes found in the chicken B complex, should clarify the extent to which genomic processes in these two groups are similar.

Materials and Methods

Cosmid isolation. Rwcos3 was screened from a genomic library [13] of a female red-winged blackbird from Washington state, using probes spanning exons 1–4 of a red-winged blackbird class II RT-PCR product [31]. Rwcos3 was grown in LB media for 16 h and isolated using a Qiagen (Valencia, CA.) maxiprep kit.

Sequencing and assembly. To prepare for sequencing, the cosmid was sonicated, ragged ends repaired using T4 DNA polymerase and free nucleotides, size fractionated on an agarose gel (2–4 kb), the fragments ligated into M13 vector and isolated as single stranded DNA using the Qiagen (Valencia, CA) 96-well M13 prep kit. We used a modified primer M13.5 (5′-TGCCTGCAGGTCGACTCTAG-3′) to obtain usable sequence within 10 bp of the cloning site. 567 random subclones were sequenced using BigDye chemistry on a 377 ABI machine. Chromatograms were aligned and assembled into contigs using the UNIX environment programs Phred [33] and Phrap (Phil Green, unpublished data). The perlscript 'phredPhrap' (David Gordon, unpublished data) coordinates the calling of the alignment programs. The resulting contigs were visualized using Consed [34] and connected to one another by designing primers at each end and sequencing appropriate subclones.

Sequence analysis. The final contig was analyzed for possible genes, exons, and CpG islands using SeqHelp [35] and associated internal modules such as Genscan [36], which identifies promoter regions by searching for TATA boxes followed by initiator signals. Microsatellites were determined using the program Sputnik (Chris Abajian, unpublished data) and retroelements were identified by comparison with sequences available in the database Repbase using Censor (37; http://www.girinst.org) and a local database using RepeatMasker (Arian Smit and Phil Green, unpublished data). Both programs screen the DNA sequence using a library of known mammalian, fish, and amphibian repeats. A custom library was constructed for screening with RepeatMasker consisting of CR1 (chicken repeat 1) elements, which are a class of LINEs thus far found only in chickens [20], and *gag* sequences recently characterized found in grouse (*Tetraoninae*) [38]. We also re-analyzed Rwcos10 with all of these methods. Rwcos3 and Rwcos10 were compared with each other using the GCG package programs "Compare" and "Dotplot," using a stringency of 33/50 (Genetics Computer Group, Madison, Wisconsin).

Polymorphism of Agph-DAB3 exon 2. We designed primers in less-conserved regions of introns 1 and 2 (DAB3_exon2F, 5′-GAACTTTGGGGGTGCTGT-3′, and DAB3_exon2R, 5′-GAAATGACTGGTCATGGACT-3′) to amplify exon 2 sequences specific to *Agph-DAB3*. Nine blackbirds were amplified and directly sequenced in both directions using the amplification primers. The sequences were aligned by Phrap, heterozygous sites were called by PolyPhred [39] and inspected manually. PolyPhred finds potential heterozygotes in the chromatograms based on a predictable reduction in peak heights when compared with homozygous sites. The program Hapinfer [40] reconstructed the haplotypes from the unphased diploid sequences. Hapinfer constructs haplotypes through a heuristic chain of inference in which homozygotes are first determined, then the remaining haplotypes are made from combinations of the remaining ambiguous sites. We used the software package Mega [41] to calculate the number of synonymous (d_S) and nonsynonymous (d_N) substitutions per site using the Jukes-Cantor method [42]. Other sequence statistics, such as Tajima's *D*, a measure of the departures of sequence patterns from neutral expectations [43], and the number of segregating (polymorphic) sites were calculated using DNAsp [44]. Mega was also used to determine the total number of substitutions per site *d* between the duplicated segments using the Kimura 2-parameter distance method [45]. The relationship $d = 2\mu T$, where μ is the substitution rate and T is the divergence time, was used to infer the times of divergence of the duplicated segments and also the divergence times among *Agph-DAB1*, *Agph-DAB2*, and *Agph-DAB3*. A silent (noncoding) mutation rate of 1×10^{-9} substitutions per site per year (μ) has been estimated for introns and exons of human DRB genes and was used in this study [32]. This mutation rate was estimated under the assumption of, and presumably in the absence of, gene conversion, which is known to elevate silent rates in *Mhc* genes [46]. Thus, this estimate will be more appropriate for blackbird regions probably not experiencing gene conversion (such as noncoding DNA, retroelements, and introns) than to those that are (such as peptide-binding exons).

Phylogenetic analysis. We analyzed phylogenetically previously published exon 2 haplotypes from blackbird genes *Agph-DAB1* [14] and *Agph-DAB2* [11], in addition to the new haplotypes from *Agph-DAB3* ($n = 6$; this study), and used a *Mhc* class II sequence from a Scrub Jay (*Apelocoma coerulescens*) [31] as an outgroup. The *Agph-DAB2* and *Agph-DAB3* alleles consist of full-length exon 2 sequences (269 bases), whereas the *Agph-DAB1* haplotypes represent the last 89 bases. We used the neighbor-joining method with the Kimura 2-parameter model as implemented in the program PAUP*4b [47] to construct the phylogenetic relationships and tested branches for significance with 2000 bootstrap replicates.

Acknowledgments

We thank Hollie Walsh, Chris Hess, Robb Brumfield, Monica Silva, Andrew Shedlock, and Holly A. Wichman for comments; Mark Reider and Dana Carrington for technical assistance with PolyPhred and Hapinfer; Todd Smith for DrawMap; and Noriko Kitamura Gasper for formatting the references and some typing. This work was supported by NSF grant DEB-9815800 to S. V. E.

RECEIVED FOR PUBLICATION DECEMBER 26, 2000;
ACCEPTED MAY 9, 2001.

References

1. Klein, J. (1986). *Natural History of the Major Histocompatiblity Complex*. Wiley, New York.
2. Hughes, A. L. (2000). *Adaptive Evolution of Genes and Genomes*. Oxford University Press, Oxford.
3 Beck, S., and Trowsdale, J. (1999). Sequence organisation of the class II region of the human *Mhc*. *Immunol. Rev.* **167:** 201–210.
4. Kulski, J. K., *et al*. (1997). The evolution of *MHC* diversity by segmental duplication and transposition of retroelements. *J. Mol. Evol.* **45:** 599–609.
5. Yamazaki, M., Tateno, Y., and Inoko, H. (1999). Genomic organization around the centromeric end of the HLA class I region: large-scale sequence analysis. *J. Mol. Evol.* **48:** 317–327.
6. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

7. Kaufman, J., *et al.* (1999). The chicken locus is a minimal-essential major histocompatibility complex. *Nature* **401:** 923–925.

8. Kaufman, J., Völk, H., and Wallny, H. -J. (1995). A "minimal essential *Mhc*" and an "unrecognized *Mhc*": two extremes in selection for polymorphism. *Immunol. Rev.* **143:** 63–88.

9. Hala, K., *et al*. (1988). Attempt to detect recombination between B-F and B-L genes within the chicken B complex by serological typing, in vitro MLR and RFLP analyses. *Immunogenetics* **28:** 433.

10. Sibley, C. G., and Ahlquist, J. E. (1990). *The Phylogeny and Classification of Birds: A Study in Molecular Evolution*. Yale Univ. Press, New Haven.

11. Edwards, S. V., Gasper, J., Garrigan, D., Martindale, D., and Koop, B. (2000). A 39-kb sequence around a blackbird *Mhc* class II gene: Ghost of selection past and songbird genome architecture. *Mol. Biol. Evol.* **17:** 1384–1395.

12. Hess, C. M., Gasper, J., Hoekstra, H., Hill, C., and Edwards, S. V. (2000). *Mhc* class II pseudogene and genomic signature of a 32-kb cosmid in the house finch (*Carpodacus mexicanus*). *Genome Res.* **10:** 613–623.

13. Edwards, S. V., Gasper, J., and Stone, M. (1998). Genomics and polymorphism of *Agph-DAB1*, an *Mhc* class II B gene in red-winged blackbirds (*Agelaius phoeniceus*). *Mol. Biol. Evol.* **15:** 236–250.

14. Garrigan, D., and Edwards, S. V. (1999). Polymorphism across an intron exon boundry in an avian *Mhc* class II B gene. *Mol. Biol. Evol.* **16:** 1599–1606.\

15. McQueen, H. A., *et al*. (1996). CpG islands of chicken are concentrated on microchromosomes. *Nat. Genet.* **12:** 321–324.

16. Odeberg, J., *et al*. (1998). Cloning and characterization of ZNF189, a novel human *Krüppel*-like zinc finger gene localized to chromosome 9q22-q31. *Genomics* **50:** 213–221.

17. Matzuk, M. M., and Bradley, A. (1992). Structure of the mouse activin receptor type II gene. *Biochem. Biophys. Res. Commun.* **185:** 404–413.

18. Zoorob, R., Béhar, G., Kroemer, G., and Auffrey, C. (1990). Organization of a functional chicken class II β gene. *Immunogenetics* **31:** 179–187.

19. Benoist, C., and Mathis, D. (1990). Regulation of major histocompatibility complex class II genes: X, Y and other letters of the alphabet. *Annu. Rev. Immunol.* **8:** 681–715.

20. Burch, J. B. E., Davis, D. L., and Haas, N. B. (1993). Chicken repeat 1 elements contain a *pol*-like open reading frame and belong to the non-long terminal repeat class of retrotransposons. *Proc. Natl. Acad. Sci. USA* **90:** 8199–8203.

21. Kaufman, J., Salomonsen, J., and Flajnik, M. (1994). Evolutionary conservation of *MHC* class I and class II molecules–different yet the same. *Semin. Immunol.* **6:** 411–424.

22. Edwards, S. V., Grahn, M., and Potts, W. K. (1995). Dynamics of *Mhc* evolution in birds and crocodilians: amplification of class II genes with degenerate primers. *Mol. Ecol.* **4:** 719–729.

23. Kaufman, J., *et al.* (1999c). Gene organisation determines evolution of function in the chicken *MHC*. *Immunol. Rev.* **167:** 101–117.

24. Shiina, T., *et al*. (1999b). Gene organization of the quail major histocompatibility complex (*MhcCoja*) class I gene region. *Immunogenetics* **49:** 384–394.

25. Fischer Lindahl, K., *et al*. (1997). H2-M3, a full service class I_B histocompatibility antigen. *Ann. Rev. Immunol.* **15:** 851–879.

26. Wittzell, H., Bernot, A., Auffray, C., and Zoorob, R. (1999). Concerted evolution of two *Mhc* class II B loci in pheasants and domestic chickens. *Mol. Bio. Evol.* **16:** 479–490.

27. Shedlock, A. M., Milinkovitch, M. C., and Okada, N. (2000). SINE evolution, missing data and the origin of whales. *Syst. Biol.* **49:** 808–817.

28. Hughes, A. L., and Hughes, M. K. (1995). Small genomes for better flyers. *Nature* **377:** 391.

29. Clarke, M. S., *et al.* (1999). Sequence scanning chicken cosmids: a methodology for genome screening. *Gene* **227:** 223–230.

30. Gongora, R. (1997). Presence of solitary exon 1 sequences in the *HLA-DR* region. *Hereditas* **127:** 47–49.

31. Edwards, S. V., Wakeland, E. K., and Potts, W. K. (1995). Contrasting histories of avian and mammalian *Mhc* genes revealed by class II B sequences from songbirds. *Proc. Natl. Acad. Sci. USA* **92:** 12200–12204.

32. Satta, Y., O'Huigin, C., Takahata, N., and Klein, J. (1993). The synonymous substitution rate at major histocompatibility complex loci in primates. *Proc. Natl. Acad. Sci. USA* **90:** 7480–7484.

33. Ewing, B., and Green, P. (1998). Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Res.* **8:** 186–194.

34. Gordon, D., Abajian, C., and Green, P. (1998). *Consed*: a graphical tool for sequence finishing. *Genome Res.* **8:** 195–202.

35. Lee, M., Lynch, E. D., and King, M. -C. (1998). SeqHelp: a program to analyze molecular sequences utilizing common computational resources. *Genome Res.* **8:** 306–312.

36. Burge, C. B., and Karlin, S. (1998). Finding genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8:** 346–354.

37. Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. (1996). CENSOR—a program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **20:** 119–122.

38. Dimcheef, D. E., Drovetski, S. V., Krishnan, M., and Mindell, D. P. (2000). Cospeciation and horizontal transmission of avian Sarcoma and Leukosis virus *gag* genes in Galliform birds. *J. Virol.* **74:** 3984–3995.

39. Nickerson, D. A., Tobe, V. O., and Taylor, S. L. (1997). PolyPhred: automating the detection and genotyping of single nucleotide substitutions using flourescence-based resequencing. *Nucleic Acids Res.* **25:** 2745–2751.

40. Clark, A. G. (1990). Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol. Biol. Evol.* **7:** 111–112.

41. Kumar, S., Tamura, K., and Nei, M. (1993). *MEGA: Molecular Evolutionary Genetic Analysis. Version 1.01*. The Pennsylvania State University, University Park.

42. Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.

43. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123:** 585–595.

44. Rozas, J., and Rozas, R. (1997). DNAsp version 2.0: a novel software package for extensive molecular population genetics analysis. *Comput. Appl. Biosci.* **13:** 307–311.

45 Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16:** 111–120.

46. Ohta, T. (1998). On the pattern of polymorphism at major histocompatibility complex loci. *J. Mol. Evol.* **46:** 633–638.

47. Swofford, D. (2000). *PAUP*4b: Phylogenetic Analysis Using Parsimony and Other Methods (Version 4b)*. Sinauer Associates, Sunderland.

48. Kelly, A., and Trowsdale, J. (1985). Complete nucleotide sequence of a functional HLA-DPβ gene and the region between the DPβ1 and DPα1 genes: comparison of the 5′ ends of HLA class II genes. *Nucleic Acids Res.* **13:** 1607–1621.

49 Auffray, C., *et al.* (1987). Structure and expression of HLA-DQα and -DXα genes: interallelic alternate splicing of the HLA-DQα gene and functional splicing of the HLA-DXα gene using a retroviral vector. *Immunogenetics* **26:** 63-73.

50. Dedrick, R. L., and Jones, P. P. (1990). Sequence elements required for activity of a murine major histocompatibility complex class II promoter bind common and cell-type specific nuclear factor. *Mol. Cell. Biol.* **10:** 593-604.

Sequences from this article have been deposited with the EMBL/GenBank Data Libraries under accession numbers AF328732–AF328737 (Agph-DAB3 haplotypes) and AF328738 (cosmid sequence).