Molecular Ecology (2002) **11**, 603–612

Single-nucleotide polymorphism characterization in species with limited available sequence information: high nucleotide diversity revealed in the avian genome

C. R. PRIMMER,* T. BORGE,† J. LINDELL‡ and G.-P. SÆTRE†

*Division of Population Biology, Department of Ecology and Systematics, University of Helsinki, PO Box 65, 00014, Helsinki, Finland, †Department of Evolutionary Biology, Evolutionary Biology Center, Uppsala University, Uppsala, Sweden, ‡Center for Biodiversity and Conservation Biology, Royal Ontario Museum, Toronto, Ontario, Canada

Abstract

As a case study for single-nucleotide polymorphism (SNP) identification in species for which little or no sequence information is available, we investigated several approaches to identifying SNPs in two passerine bird species: pied and collared flycatchers (*Ficedula hypoleuca* and *F. albicollis*). All approaches were successful in identifying sequence polymorphism and over 50 candidate SNPs per species were identified from ≈ 9.1 kb of sequence. In addition, 17 sites were identified in which the frequency of alternative bases differed by > 50% between species (termed interspecific SNPs). Interestingly, polymorphism of microsatellite/intron loci in the source species appeared to be a positive predictor of nucleotide diversity in homologous flycatcher sequences. The overall nucleotide diversity of flycatchers was $2.3-2.7 \times 10^{-3}$, which is ≈ 3–6 times higher than observed in recent studies of human SNPs. Higher nucleotide diversity in the avian genome could be due to the relatively older age of flycatcher populations, compared with humans, and/or a higher long-term effective population size.

Keywords: bird, flycatcher, gene, genomics, intron, single nucleotide polymorphism

Received 9 June 2001; revision received 3 December 2001; accepted 3 December 2001

Introduction

Single-nucleotide polymorphisms (SNPs) can be broadly defined as any single base substitution/indel in the genome of an individual. They have been found to occur every 100–300 bp in humans, most are bi-allelic and inherited in a Mendelian fashion (see Brown 1999). Owing in part to their high frequency, also in coding DNA sequences, and the possibility for using highly automated analysis systems (reviewed in Landegren *et al*. 1998), the utilization of SNPs as genetic markers has received much attention recently in human genetic studies of, for example, gene mapping (Wang *et al*. 1998; The International SNP Map Working Group 2001) and human evolution (Cargill *et al*. 1999; Hacia *et al*. 1999). Similarly, in other species for which large-scale genome projects are underway, SNPs are being identified at a rapid rate (Lindblad-Toh *et al*. 2000; Marklund *et al*. 2000; Hoskins *et al*. 2001). In addition to recent advances in DNA sequencing and microarray technologies (Meldrum 2000a,b), one of the key factors facilitating the rapid characterization of SNPs in well-studied species has been the public availability of large amounts of overlapping sequence data, thus enabling the identification of sequence polymorphisms with a reduced amount of laboratory work: so-called 'data-mining' (e.g. Buetow *et al*. 1999).

Despite numerous potential applications, identification of SNP markers in free-living vertebrate species has been almost nonexistent. This is likely because of the limited amount of nuclear sequence data available for most species, which in turn limits the possible sequences that can be investigated. This is particularly so for avian species, in which a recent examination of the GenBank sequence database (Benson *et al*. 2000) revealed that a mere 0.4% of vertebrate sequence entries are derived from avian species and two-thirds of these sequences are from a single species, the domestic chicken *Gallus gallus*.

The aim of this study was to characterize SNPs in two passerine bird species, the pied flycatcher (*Ficedula hypoleuca*) and the collared flycatcher (*F. albicollis*). The *Ficedula* flycatchers have been the focus of extensive ecological

Correspondence: C. Primmer. Fax: +358-9191-57694; E-mail: Craig.Primmer@helsinki.fi

and evolutionary research (reviewed by Lundberg & Alatalo 1992) and potential applications of SNP markers include analysis of postglacial colonization routes of the species and studying the influence of gene sequence variation in the processes of sexual selection and life history evolution. Of particular interest, the *Ficedula* flycatcher species complex has become an important model system for studies of speciation and hybridization (e.g. Alatalo *et al*. 1994; Sætre *et al*. 1997, 2001). Thus, an important application of SNPs would be to investigate the genetics of speciation, including introgression of alleles in hybrid zones and identification of candidate genes affecting hybrid fitness. Prior to this study, a total of just 14 nuclear-derived sequences were available from the two flycatcher species, the majority representing microsatellite loci. We were therefore interested in exploring potential alternatives for identifying nuclear SNPs in the *Ficedula* flycatchers, with a view to also providing guidelines for SNP identification in other species for which available sequence information is extremely limited.

The first challenge was to identify flycatcher DNA sequences from which SNPs could be subsequently screened. Two alternative approaches were tested for flycatcher sequence characterization. The first strategy considered was the utilization of polymerase chain reaction (PCR) primers designed using sequences obtained from other avian species, and subsequent PCR amplification of homologous flycatcher sequences. Two main sources of DNA sequences were utilized in this approach, the first was an adaptation of the comparative anchor-tagged sequence (CATS) method, which has been utilized extensively for comparative genome mapping in mammals (e.g. O'Brien *et al*. 1993; Lyons *et al*. 1997). Using this approach, PCR primer pairs were designed to span intronic regions based on the sequence of a related species (primarily the domestic chicken in this study), but the primers themselves were designed in exonic sequences in order to increase the evolutionary conservation of the sequences. Avian introns have previously been shown to be an abundant source of sequence variation as revealed by single-stranded conformation polymorphism (SSCP) analyses (Friesen *et al*. 1997). A second source of evolutionarily conserved avian sequences was also utilized, namely microsatellite sequences identified from other passerine bird species. Although exhibiting a much higher evolutionary rate (and therefore lower rate of sequence conservation) than exon sequences, it has been shown that a certain proportion of avian microsatellites amplify in related species (e.g. Primmer *et al*. 1996), and over 400 microsatellites from Passeriformes, the order to which the pied and collared flycatchers belong, are currently available in GenBank. As the aim was to identify single-base differences in the flanking sequences of homologous products, rather than identification of length polymorphisms in the microsatellite repeat region, length monomorphic sequences were also considered as a potential source for SNPs.

The second approach investigated was sequencing of random clones obtained from pied and collared flycatcher genomic DNA libraries. This approach has the obvious advantage of potentially identifying large numbers of sequences from the focal species from which highly homologous PCR primers can be designed. However, one possible disadvantage of this approach is the fact that repetitive sequences may be over-represented.

Materials and methods

Candidate sequence identification of flycatcher nuclear DNA sequences

Candidate avian intron sequences were extracted from GenBank using the criteria that they contained intronic regions of ≈ 400–600 bp in length, i.e. to enable single read sequencing of nearly the entire PCR products, and that there was sufficient exon sequence on either side of the intron for primer design. Candidate sequences were then subjected to a BLAST (Altschul *et al*. 1997) homology search. Sequences which were members of closely related gene families, or for which pseudogenes had been identified, were excluded. PCR primers were then designed using the program PRIMER 3 (Rozen & Skaletsky 1998), with the aim of amplifying a PCR product length of 400–600 bp. In total, 41 primer pairs were designed from a total of 636 sequence entries which met the above criteria. In addition, eight primer pairs which had been previously shown to amplify polymorphic intronic sequences in a nonpasserine bird, the marbled murrelet, *Brachyramphus marmoratus* (Friesen *et al*. 1997, 1999), were tested. Candidate microsatellite sequences were limited to those identified in species from the same order as flycatchers, i.e. passerine species. In total, 34 microsatellite loci were tested. In all cases, primer sequences reported for the source species were utilized. In addition, PCR primers were designed based on sequences of 14 random clones from a pied flycatcher genomic DNA library and one clone from a collared flycatcher DNA library (see below).

PCR amplification and sequencing

Total genomic DNA was isolated from blood samples, according to the protocol quoted in Haavie *et al*. (2000), from pied flycatchers sampled near Oslo, Norway (n = 10), and near Madrid, Spain (n = 10) and collared flycatchers from Breclav, Czech Republic (n = 20). PCR reactions for gene sequences were performed using either MJ Research PTC 100/PTC 200 or Eppendorf gradient thermal cyclers in 20 μL volumes with ≈ 100 ng of DNA, 0.2 units of Biotaq DNA polymerase (Bioline), 250 μM of each dNTP and PCR buffer containing 1.5 mM $MgCl_2$, 16.0 mM $[(NH4)_2\ SO_4]$,

67.0 mM Tris–HCl (pH 8.8), 0.01% Tween-20. In general, PCR amplification was first attempted using a 'touchdown' PCR protocol as described in Koskinen & Primmer (1999) except that a 45 s 72 °C extension was used. Loci for which potentially homologous PCR products were observed following agarose gel electrophoresis were then subjected to more extensive PCR optimization using the following PCR profile: 94 °C for 2 min, followed by 35 cycles of 95 °C for 30 s, one of various annealing temperatures for 30 s, 72 °C for 45 s, with a final 72 °C extension of 5 min. Adjustment of $MgCl_2$ concentration was not attempted in the optimization phase as it has been shown to be less important for obtaining amplification of specific products than the annealing temperature (Morin *et al*. 1998).

Detailed methodology for sequence characterization from flycatcher genomic DNA libraries is given in Sætre *et al*. (2001). Fourteen clones from a pied flycatcher DNA library and one from a collared flycatcher DNA library were prepared for sequencing using the QIAprep Spin Miniprep Kit (Qiagen). Sequencing reactions were performed as listed below and PCR primers to amplify flycatcher genomic DNA sequences were designed using primer design software (Oligo) based on the sequences of 13 clones. Primer sequences, PCR annealing temperatures and additional details for all loci further investigated are listed in Appendix 1. Details of loci not further investigated are available on request.

In order to screen for sequence polymorphism, PCR products were amplified from between two and eight individuals per species (average 3.7; Appendix 1). PCR procedures for amplifying conserved gene and microsatellite sequences were as described above, and PCR procedures for amplification of random clones are given in Saetre *et al*. (2001). Unincorporated primers and dNTPs were removed from PCR products prior to sequencing either using exonuclease 1-shrimp alkase phosphatase treatment (Amersham Pharmacia Biotech) or by using QIAquick PCR purification columns (Qiagen). Sequences were obtained using BigDye dye-terminator (Version 1) chemistry (Applied Biosystems) following the manufacturer's recommendations and using one of the primers used for PCR.

SNP identification from sequence data

Sequences were base-called and aligned using either Sequence Analysis (Applied Biosystems) and Sequencher (Gene Codes Corp.) or the 'SNP pipeline' (Buetow *et al*. 1999; available from http://lpgws.nci.nih.gov:82/perl/snp/snp_cgi.pl) which applies the Phred/Phrap/PolyPhred series of base-calling, alignment and SNP identification programs (Nickerson *et al*. 1997; Ewing & Green 1998; Ewing *et al*. 1998). Homologous sites from each locus alignment that appeared to exhibit sequence variation, either in heterozygous or homozygous forms, were evaluated by manual inspection. Previous studies have shown that a small proportion of candidate polymorphisms may be false positives because of, for example, sequencing artefacts (e.g. Ewing & Green 1998; Wang *et al*. 1998). In order to minimize such cases in our data-set, we used the following assessment criteria: (i) whether the base had been sequenced in both directions; (ii) whether a similar sequence was observed in more than one individual; and (iii) whether the potential polymorphism occurred in a region of generally high sequence quality (phred score ≥ 20). High-quality sequences in both directions were obtained for at least one individual possessing the rarer nucleotide variant for 80% of polymorphic sites included in the study. The rarer nucleotide variant of the remaining 20% of sites was either observed in the sequences of two or more individuals, or in one individual in a region of high sequence quality. Low-quality single-read sequence regions were excluded from all analyses.

In addition to identifying candidate SNPs within each species, cases of base frequency differences between species (interspecific SNPs) were also noted. An interspecific SNP was defined as any case in which the frequency of alternative alleles at homologous sites differed by > 50% between species. Nucleotide polymorphism for a subset of SNPs was confirmed by additional analysis methods (see below), however, given that it is still possible that a proportion of the apparent nonconfirmed polymorphisms may be the result of sequencing artefacts (e.g. Wang *et al*. 1998) we use the phrase 'candidate SNPs'.

Nucleotide diversity was estimated, based on the normalized number of polymorphic sites, which corrects for differing sequence lengths and variation in the number of gene copies analysed (Watterson 1975; Nei 1987), by using the formula θ (also called k) $= K/L * [1^{-1} + 2^{-1} + 3^{-1} + ... + (n-1)^{-1}]$, where K is the number of polymorphic sites, L is the sequence length (in bp) and n is the number of chromosomes screened. Tests for correlation between diversity levels of particular loci in the source species and the nucleotide diversity observed in flycatchers were performed by comparing the published expected heterozygosity in the source species with θ-values estimated from flycatcher sequences (data from both species pooled).

Results

Of 98 primer pairs tested, high-quality sequences were obtained for 28 loci, including 18% of the passerine microsatellites, 20% of gene sequences extracted from GenBank and 53% of the sequences from the flycatcher genomic DNA libraries (Table 1). In addition, sequences were obtained for six of eight avian introns previously shown to be conserved between two nonpasserine bird species (Table 1). Sequence polymorphism was observed either within or between species for all but five of the 28

Table 1 Summary of sequence generation and candidate SNP identification success rates within and between pied and collared flycatchers

Sequence type	No. loci tested	No. loci sequenced	Av. bp/locus	Proportion of sequenced loci revealing candidate SNPs (Av. number of candidate SNPs per locus)		
				Pied	Collared	Interspecific*
GenBank avian introns	41	8	473	0.75 (2.1)	0.88 (3.0)	0.50 (0.88)
'Conserved' avian introns†	8	6	331	0.83 (1.8)	0.83 (2.2)	0.50 (0.67)
Microsatellites	34	6‡	117	0.50 (1.3)	0.50 (1.7)	0 (0)
Random clones	15	8	331	0.63 (2.0)	0.63 (1.8)	0.38 (0.75)

*Interspecific candidate SNPs were defined as cases where the frequency of alternative alleles differed by > 50% between species.
†Primer sequences taken from Friesen *et al*. (1997, 1999).
‡An additional two loci for which a clear product was obtained were excluded due to their short length.

sequenced loci with the proportion of sequences revealing polymorphism within either species ranging from 50% of microsatellite loci to over 80% of intron sequences (Table 1). In addition to following strict sequence evaluation criteria so as to limit the number of false-positive polymorphisms in the dataset (see Materials and methods), we also confirmed the presence of sequence polymorphism in a subset of the loci using other analysis methods. Mendelian inheritance of an interspecific sequence polymorphism at the *LAMA* (Lamin A) locus was confirmed by analysing PCR products obtained from a one-generation hybrid family using restriction fragment length polymorphism (RFLP) analysis (Fig. 1). In addition, the occurrence of heterozygous bases in PCR products at the *TGFB2* and *RDPSN* loci was confirmed using denotaring high-performance liquid chromatography (DHPLC) analyses (G-PS unpublished). Finally, the number of SSCP genotypes observed at the *RDPSN*, *Fh33B* and *Fh69* loci were in complete agreement with sequence data (G-PS unpublished).

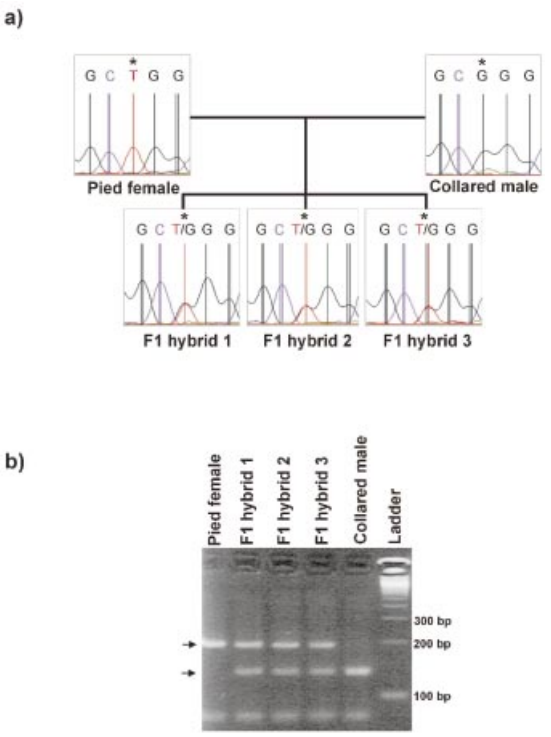

Fig. 1 Mendelian inheritance in an aviary-reared hybrid family of an interspecific single nucleotide polymorphism identified at the *LAMA* locus as observed by (a) sequence analysis (the polymorphic base is indicated by an asterix); and (b) cleavage of *LAMA* PCR products with the *Sdu*I restriction enzyme (polymorphic bands indicated by arrows).

Intraspecific SNP characterization

The total number of base pairs screened per species was 9082 and 9164 for pied and collared flycatchers, respectively, from which 52 and 61 candidate intraspecific SNPs were identified (Table 2). This represents an average frequency of one SNP per 175 and 150 bp for pied and collared flycatchers, respectively. The polymorphic sites included a small number of indels for each species (Table 2), giving an estimated average frequency of one indel per 2271 and 4582 bp in pied and collared flycatchers, respectively. Nucleotide diversity of the microsatellite sequences was around double that of other analysed sequence classes ($4.5–5.7 \times 10^{-3}$ vs. $1.8–2.9 \times 10^{-3}$). Nucleotide diversity was relatively similar between species with overall θ-values of pied and collared flycatchers being 2.3×10^{-3} (locus range $0–20 \times 10^{-3}$) and 2.7×10^{-3} (locus range $0–19 \times 10^{-3}$), respectively (Table 2; Fig. 2).

Correlation between source species genetic diversity and flycatcher nucleotide diversity

As some of the sequences had previously been shown to exhibit length polymorphism (microsatellite loci *BMC4*, *HrU3*, *Mcyμ5*, *Zl35*, *Zl38*) or sequence polymorphism (intron sequences from the *MPLPR*, *ODC*, *LRPP40*

Table 2 Details of candidate single-nucleotide polymorphisms identified within and between pied and collared flycatchers from different subclasses of avian DNA sequence. Details of candidate gene sequence SNPs have been submitted to the database dbSNP under Accession nos ss4323248–ss4323312

	Pied flycatcher				Collared flycatcher				Inter-specific*		
Locus	bp screened	No. SNPs	SNP freq. (bp^{-1})	$\theta \times 10^3$	bp screened	No. SNPs	SNP freq. (bp^{-1})	$\theta \times 10^3$ bp	bp screened	No. SNPs	SNP freq. (bp^{-1})
Gene sequences											
CEPUS	590	4†	148	6.4	568	2	284	3.5	568	2†	284
HNFAL	106	1†	106	3.6	116	1†	116	4.7	106	0	—
ACL	487	0	—	0	487	1	487	0.8	485	0	—
RDPSN	920	1	920	0.4	933	4	233	1.7	920	3	307
TGFB2	592	2	296	1.1	592	5	118	2.8	592	1	592
A-B crystallin	117	0	—	0	117	0	—	0	117	0	—
Tropomyosin	489	6	82	4.7	489	5	98	3.6	489	0	—
ALASY	480	3	160	3.4	480	6	80	6.8	480	1	480
GADPH	298	1	298	1.1	298	4	75	4.0	298	0	—
LAMA	213	0	—	0	213	0	—	0	213	1	213
AETC	256	1	256	1.5	256	1	256	1.7	256	0	—
MPLPR	301	2	151	2.2	226	2	113	3.4	226	0	—
LRPP40	349	6†	58	6.6	349	4	87	4.4	349	2†	175
ODC	555	1	555	0.7	656	2	328	1.2	555	1	555
Sub-total	5753	28	205	1.8	5780	37	156	2.5	5656	11	514
Microsatellites											
BMC4	94	0	—	0	98	0	—	0	94	0	—
HrU3	167	5	33	20	167	8	21	19	167	0	—
Mcyμ5	129	2	65	6.0	151	0	—	0	129	0	—
Pocc8	161	0	—	0	161	1	161	2.4	161	0	—
ZL35	73	1	73	4.8	73	1	73	4.8	73	0	—
ZL38	68	0	—	0	68	0	—	0	68	0	—
Sub-total	692	8	87	4.5	718	10	72	5.7	692	0	< 692
Random clones											
Fa45B	213	0	—	0	213	4	53	6.2	213	1	213
Fh33	375	0	—	0	375	0	—	0	375	0	—
Fh33B	146	3	49	7.9	146	1	146	3.7	146	0	—
Fh45	372	3	124	3.5	401	3	134	3.3	372	4	93
Fh49	311	0	—	0	311	0	—	0	311	0	—
Fh61	435	3	145	3.8	435	5	87	4.1	435	1	435
Fh63	409	6†	68	5.7	409	1†	409	1.1	409	0	—
Fh69	376	1	376	1.2	376	0	—	0	376	0	—
Sub-total	2637	16	165	2.9	2666	14	190	2.3	2637	6	440
Overall	**9082**	**52**	**175**	**2.3**	**9164**	**61**	**150**	**2.7**	**8985**	**17**	**529**

*An interspecific candidate SNP is defined as any case where the frequency of alternative alleles differs by > 50% between species.
†Includes one indel polymorphism.

genes) in other avian species, it was possible to examine if there was any apparent correlation between the level of polymorphism in the source species, and the sequence diversity of the homologous flycatcher sequences. Interestingly, polymorphism level in the source species (as estimated by expected heterozygosity; see Appendix 1 for references) was found to be positively correlated with the sequence diversity (as estimated by θ) of the homologous loci in flycatchers (Spearman rank correlation r_s = 0.73, Z = 2.08, P = 0.037; Fig. 3).

Interspecific SNP characterization

Allele frequencies differed markedly between species at 17 homologous sites indicating interspecific SNPs to occur, on average, once every 529 bp, suggesting that the genomes of pied and collared flycatchers differ at $\geq 2.2 \times 10^6$ bp (≈0.2%), assuming an avian genome size of 1.2×10^9 bp (Cavalier-Smith 1985). Interspecific SNPs were found most frequently from the random flycatcher clones (one per 440 bp), whereas no interspecific SNPs were

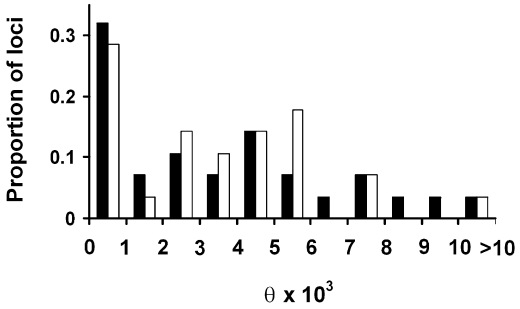

Fig. 2 Frequency distribution of nucleotide diversity (θ), based on the normalized number of polymorphic sites, observed from the sequences of 18 loci investigated in pied and collared flycatchers (black and white bars, respectively).

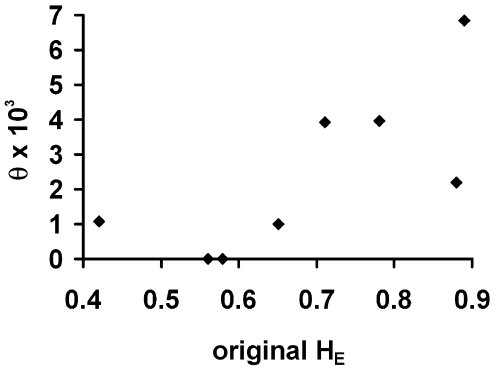

Fig. 3 Positive association between expected heteorzygosity (H_E) of selected loci in the source species, and nucleotide diversity (θ) in pied and collared flycatchers (Spearman rank correlation $r_s = 0.73$, $Z = 2.08$, $P = 0.037$).

observed in the 692 bp of microsatellite sequence compared (Table 2).

Discussion

Of the various sequence classes tested, the highest amplification success rate in flycatchers was, not surprisingly, achieved for intronic sequences previously shown to be conserved in other birds (75%) and for sequences obtained from flycatcher genomic DNA libraries (57%). The amplification success rate of primer pairs designed in exon sequences to amplify avian introns, and from passerine microsatellite loci, was lower (20 and 18%, respectively; Table 2). However, given that, at the time of writing, >1000 avian intron/microsatellite sequences fitting our original selection criteria (see Materials and methods) were available from GenBank, they should provide a valuable source of sequences for identifying additional SNP loci in numerous bird species. As the number of sequence entries for other animal lineages, such as mammals and fish, is many times greater that available for birds, it is likely that a similar SNP identification approach could also be utilized in other species groups.

All sequence classes tested proved to be valuable sources for identifying intraspecific SNPs in flycatchers. The highest nucleotide diversity was observed in microsatellite sequences suggesting that despite the absence of tandem repeat regions of any notable length in the flycatcher homologue sequences (data not shown) these genomic regions are still subjected to higher than average mutation rates. However, given that primer sequences available for most microsatellite loci produce relatively short PCR fragments (the average microsatellite sequence length in this study was 117 bp), higher nucleotide diversity does not necessarily result in a net gain of the number of polymorphic sites per locus (Table 1). Also, the concentration of most of the microsatellite SNPs within just two of the six loci examined may have implications depending on whether tightly linked or unlinked SNPs are desired (see Nielsen 2000). Microsatellites also appeared to exhibit a lower frequency of interspecific SNP differences (Tables 1 and 2), although examination of a greater amount of microsatellite sequence data would be required to confirm this result. The observed overall rate of interspecific SNPs (one per 529 bp) suggests that the sequence identification approach described here will provide a valuable source of markers for future studies of hybridization between these flycatcher species.

Interestingly, the diversity levels of loci described previously in other avian species (expected heterozygosities of microsatellite repeat length polymorphisms and of SSCP intronic sequences) appeared to be an indicator for the level of sequence variation expected in homologous flycatcher sequences (Fig. 3). This indicates that the relative mutation rates of particular genomic regions may be conserved between species. From a practical viewpoint, this correlation could be utilized for more efficient SNP identification when diversity levels have been characterized in other species.

High nucleotide diversity in the avian genome

The overall nucleotide diversity observed in pied and collared flycatchers sampled from just one or two populations ($2.3–2.7 \times 10^{-3}$) is around six times higher than that observed in a survey of human SNPs from European samples (Wang *et al*. 1998) and over three times higher than observed from SNP analysis of an ethnically diverse sample set (The International SNP Map Working Group 2001). The higher nucleotide diversity in flycatchers relative to humans could be due to several factors. It is estimated that the age of modern human populations is ≈ 200 000 years old (Pääbo 1999) and that the effective population size is ≈ 10 000 (e.g. Halushka *et al*. 1999). A higher level of sequence diversity could have been

maintained in flycatcher populations either if the average allele age is older than that in humans or because of a higher long-term effective population size. Although based on less data than human estimates (see Pääbo 1999), mitochondrial DNA data suggest that the age of pied and collared flycatcher populations may be ≈ 1.5 million years (Saetre *et al*. 2001), thereby allowing substantially more time for mutations to have accumulated in flycatchers than in humans. It is likely that the size of flycatcher populations was reduced during glacial periods because of the limiting distribution of appropriate habitat (Tegelström & Gelter 1990), however, the extent of this population reduction is difficult to estimate. It is possible, however, that such habitat reduction and fragmentation may have had less effect on species with the ability to fly because of their presumed greater dispersal abilities, therefore enabling a relatively greater proportion of sequence diversity to be retained by, for example, avian species during these periods.

Acknowledgements

Thanks to Jodie Painter and an anonymous referee for comments on an earlier version of the manuscript, and to James Kijas for helpful discussions. Leena Laaksonen provided excellent technical assistance. Part of the work presented in this study was completed while TB was visiting the Department of Ecology and Systematics, University of Helsinki. Financial support was received from the Academy of Finland and University of Helsinki (to CRP project # 172964; 2105027) and the Norwegian Research Council and the Swedish Natural Sciences Research Council (to G-PS).

References

Alatalo RV, Gustafsson L, Lundberg A (1994) Male coloration and species recognition in sympatric flycatchers. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **256**, 113–118.

Altschul SF, Madden TL, Schäffer AA *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Bensch S, Price T, Kohn J (1997) Isolation and characterization of microsatellite loci in a *Phylloscopus* warbler. *Molecular Ecology*, **6**, 91–92.

Benson DA, Karsch-Mizrachi I, Lipman DJ *et al*. (2000) GenBank. *Nucleic Acids Research*, **28**, 15–18.

Brown TA (1999) *Genomes*. BIOS Scientific Publishers, Oxford.

Buetow KH, Edmonson MN, Cassidy AB (1999) Reliable identification of large numbers of candidate SNPs from public EST data. *Nature Genetics*, **21**, 323–325.

Burt DW, Paton IR (1991) Molecular cloning and primary structure of the chicken transforming growth factor-beta2 gene. *DNA and Cell Biology*, **10**, 723–734.

Cargill M, Altshuler D, Ireland J *et al*. (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genetics*, **22**, 231–238.

Caspers GJ, Uit de Weerd D, Wattel J, de Jong WW (1997) Alpha-crystallin sequences support a galliform/anseriform clade. *Molecular Phylogenetics and Evolution*, **7**, 185–188.

Cavalier-Smith T (1985) Eukaryotic gene numbers, non-coding DNA and genome size. In: *The Evolution of Genome Size* (ed. Cavalier-Smith T), pp. 69–103. John Wiley, New York.

Clausse N, Jackers P, Jares P *et al*. (1996) Identification of the active gene coding for the metastasis-associated 37LRP/p40 multifunctional protein. *DNA and Cell Biology*, **15**, 1009–1023.

Degnan SM, Robertson BC, Clegg SM, Moritz CC (1999) Microsatellite primers for studies of gene flow and mating systems in white-eyes (*Zosterops*). *Molecular Ecology*, **8**, 159–160.

Double MC, Dawson D, Burke T, Cockburn A (1997) Finding the fathers in the least faithful bird: a microsatellite-based genotyping system for the superb fairy-wren *Malurus cyaneus*. *Molecular Ecology*, **6**, 691–693.

Ewing B, Green P (1998) Basecalling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186–194.

Ewing B, Hillier L, Wendl M, Green P (1998) Basecalling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research*, **8**, 175–185.

Friesen VL, Congdon BC, Kidd MG, Birt TP (1999) Polymerase chain reaction (PCR) primers for the amplification of five nuclear introns in vertebrates. *Molecular Ecology*, **8**, 2147–2149.

Friesen VL, Congdon BC, Walsh HE, Birt TP (1997) Intron variation in marbled murrelets detected using analyses of single-stranded conformational polymorphisms. *Molecular Ecology*, **6**, 1047–1058.

Haavie J, Saetre G-P, Moum T (2000) Discrepancies in population differentiation at microsatellites, mitochondrial DNA and plumage colour in the pied flycatcher — inferring evolutionary processes. *Molecular Ecology*, **9**, 1137–1148.

Hacia JG, Fan JB, Ryder O *et al*. (1999) Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics*, **22**, 164–167.

Halushka MK, Fan J-B, Bentley K *et al*. (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genetics*, **22**, 239–247.

Horlein A, Grajer KH, Igo-Kemenes T (1993) Genomic structure of the POU-related hepatic transcription factor HNF-1 alpha. *Biological Chemistry Hoppe-Seyler*, **374**, 419–425.

Hoskins RA, Phan AC, Naeemuddin M *et al*. (2001) Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*. *Genome Research*, **11**, 1100–1113.

Johnson R, Bulfield G (1992) Molecular cloning and sequence analysis of a chicken ornithine decarboxylase cDNA. *Animal Genetics*, **23**, 403–409.

Kim DS, Rhew TH, Moss DJ, Kim JY (1999) cDNA cloning of the CEPUS a secreted type of neural glycoprotein belonging to the immunoglobulin-like opioid binding cell adhesion molecule (OBCAM) subfamily. *Molecules and Cells*, **9**, 270–276.

Koskinen MT, Primmer CR (1999) Cross-species amplification of salmonid microsatellites which reveal polymorphism in European and Arctic grayling, Salmonidae: *Thymallus* spp. *Hereditas*, **131**, 171–176.

Landegren U, Nilsson M, Kwok P-Y (1998) Reading bits of genetic information: methods for single nucleotide polymorphism analysis. *Genome Research*, **8**, 769–776.

Lemonnier M, Balvay L, Mouly V, Libri D, Fiszman MY (1991) The chicken gene encoding the alpha isoform of tropomyosin of fast-twitch muscle fibers: organization expression and identification of the major proteins synthesized. *Gene*, **107**, 229–240.

Lindblad-Toh K, Winchester E, Daly MJ *et al*. (2000) Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genetics*, **24**, 381–386.

Lundberg A, Alatalo RV (1992) *The Pied Flycatcher*. T & AD Poyser, London.

Lyons LA, Laughlin TF, Copeland NG *etal.* (1997) Comparative anchor tagged sequences (CATS) for integrative mapping of mammalian genomes. *Nature Genetics*, **15**, 47–56.

Maguire DJ, Day AR, Borthwick IA *et al.* (1986) Nucleotide sequence of the chicken 5-aminolevulinate synthase gene. *Nucleic Acids Research*, **14**, 1379–1391.

Marklund S, Tuggle CK, Rothschild MF (2000) Mapping of the *CYP1A1*, *SSTR1* and *TTF1* genes to pig chromosome 7q refines the porcine–human comparative map. *Animal Genetics*, **31**, 318–321.

Meldrum D (2000a) Automation for genomics, part one: preparation for sequencing. *Genome Research*, **10**, 1081–1092.

Meldrum D (2000b) Automation for genomics, part two: sequencers, microarrays, and future trends. *Genome Research*, **10**, 1288–1303.

Morin PA, Mahboubi P, Wedel S, Rogers J (1998) Rapid screening and comparison of human microsatellite markers in baboons: allele size is conserved, but allele number is not. *Genomics*, **53**, 12–20.

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Nickerson DA, Tobe VO, Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research*, **25**, 2745–2751.

Nielsen R (2000) Estimation of population parameters and recombination rates from single-nucleotide polymorphisms. *Genetics*, **154**, 931–942.

O'Brien SJ, Womack JE, Lyons LA *et al.* (1993) Anchored reference loci for comparative genome mapping in mammals. *Nature Genetics*, **3**, 103–112.

Pääbo S (1999) Human evolution. *Trends in Genetics*, **15**, M13–M16.

Painter J, Crozier RH, Crozier YC, Clarke MF (1997) Characterization of microsatellite loci for a co-operatively breeding honeyeater. *Molecular Ecology*, **6**, 1103–1105.

Peter M, Kitten GT, Lehner CF *et al.* (1989) Cloning and sequencing of cDNA clones encoding chicken lamins A and B1 and comparison of the primary structures of vertebrate A- and B-type lamins. *Journal of Molecular Biology*, **208**, 393–404.

Primmer CR, Møller AP, Ellegren H (1995) Resolving genetic relationships with microsatellite markers: a parentage testing system for the swallow, *Hirundo rustica*. *Molecular Ecology*, **4**, 493–498.

Primmer CR, Møller AP, Ellegren H (1996) A wide-range survey of cross-species microsatellite amplification in birds. *Molecular Ecology*, **5**, 365–378.

Rozen S, Skaletsky HJ (1998) *Primer 3. Code available at* http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Sætre G-P, Borge T, Lindell J *et al.* (2001) Speciation, introgressive hybridisation and non-linear rate of molecular evolution in flycatchers. *Molecular Ecology*, **10**, 737–749.

Sætre G-P, Moum T, Bures S *et al.* (1997) A sexually selected character displacement in flycatchers reinforces premating isolation. *Nature*, **387**, 589–592.

Schliess F, Stoffel W (1991) Evolution of the myelin integral membrane proteins of the central nervous system. *Biological Chemistry Hoppe-Seyler*, **372**, 865–874.

Stone EM, Rothblum KN, Alevy MC, Kuo TM, Schwartz RJ (1985) Complete sequence of the chicken glyceraldehyde-3-phosphate dehydrogenase gene. *Proceedings of the National Academy of Sciences of the USA*, **82**, 1628–1632.

Takao M, Yasui A, Tokunaga F (1988) Isolation and sequence determination of the chicken rhodopsin gene. *Vision Research*, **28**, 471–480.

Tegelström H, Gelter HP (1990) Haldane's rule and sex biased gene flow between two hybridizing flycatcher species (*Ficedula albicollis* and *F. hypoleuca*, Aves: Muscicapidae). *Evolution*, **44**, 2012–2021.

The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, 928–933.

Wang DG, Fan J-B, Siao C-J *etal.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–1082.

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, **7**, 256–276.

Wistow GJ, Lietman T, Williams LA *et al.* (1988) Tau-crystallin/alpha-enolase: one gene encodes both an enzyme and a lens structural protein. *Journal of Cell Biology*, **107**, 2729–2736.

Craig Primmer's research group is interested in applying molecular genetic markers to investigate topics of ecological and evolutionary interest in nonmammalian vertebrates, as well as investigating the molecular evolution of the markers themselves. Glenn-Peter Sætre and his PhD student, Thomas Borge, are mainly investigating processes of speciation and hybrid zone dynamics, using flycatchers as model organisms. Johan Lindell completed his MSc thesis under the supervision of Sætre and Hans Ellegren.

Appendix 1 Details and analysis methods of sequences derived from (a) conserved avian intron sequences (b) avian microsatellite loci and (c) random clones isolated from flycatcher genomic DNA libraries. Sequences from this study have been submitted to GenBank under the Accession nos AJ299592-98, AJ299604-05, AJ299612-17, AJ299624-33, AJ299640-46, AJ299651-57, AJ299664-68, AJ299675-78, AJ299691-94, AJ299698-706, AF454198–282 and AY069952.

(a)	Source species						Chromosomes screened	
Locus	Accession No.	Species	Ref.*	Primer locations	Primer sequences (5′–3′)	Annealing temp. (°C)	Pied	Collared
CEPUS	AJ225897	*Gallus gallus*	1	exon 1 exon 2	F: CGAGTCAAAGTCACCGTCAA R: CTCTTCGCATCCGAGATGTA	64	6	6
HNFAL	X67690	*G. gallus*	2	exon 2 exon 3	F: GCAGCCCTCTACACCTGGTA R: CAATATCCCCTGACCAGCAT	56	8	4
ACL	AJ245665	*G. gallus*	3	exon 16 exon 17	F: GCTCTGCTTATGACAGCACT R: CAGCAATAATGGCAATGGTG	56	8	8
RDPSN	D00702	*G. gallus*	4	exon 1 exon 2	F: TGCTACATCGAGGGCTTCTT R: CGAGTGACCAGAGAGCGATT	56	10	8
TGFB2	X59080	*G. gallus*	5	exon 5 exon 6	F: GAAGCGTGCTCTAGATGCTG R: AGGCAGCAATTATCCTGCAC	65	14	12
A-B crystallin†	X96595	*Turdus merula*	6	N/A N/A	F: CTGATCCGCAGACCTTTCTT R: TAGCCAACTGGGCATCCGC	59	4	4
Tropomyosin	X57994	*G. gallus*	7	exon 6a exon 6b	F: AATGGCTGCAGAGGATAA R: TCCTCTTCAAGCTCAGCACA	60	8	10
ALASY2	X03627	*G. gallus*	8	exon 8 exon 9	F: ATTGCCCGAGTCACATCATT R: GGCTCATCAGCTTGTCAGAC	64	4	4
GADPH‡	M11213	*G. gallus*	9	exon 11 exon 12	F: ACCTTTCATGCGGGTGCTGGCATTGC R: CATCAAGTCCACAACACGGTTGCTGTA	64	12	16
LAMA‡	X16879	*G. gallus*	10	exon 3 exon 4	F: CCAAGAAGCAGCTGCAGGATGAGATGC R: CTGCCGCCCGTTGTCGATCTCCACCAG	69	10	12
AETC‡	M55140	*Anas platyrhynchos*	11	exon 8 exon 9	F: TGGACTTCAAATCCCCCGATGATCCCAGC R: CCAGGCACCCCAGTCTACCTGGTCAAA	60	8	6
MPLPR§	X61661	*G. gallus*	12	exon 4 exon 5	F: TACATCTACTTTAACACCTGGACCACCTG R: TTGCAGATGGAGAGCAGGTTGGAGCC	65	12	8
LRPP40§	X94368	*G. gallus*	13	exon 5 exon 6	F: GGGCCTGATGTGGTGGATGCTGGC R: GCTTTCTCAGCAGCAGCCTGCTC	67	8	8
ODC§	X64710	*G. gallus*	14	exon 6 exon 8	F: GACTCCAAAGCAGTTTGTCGTCTCAGTGT R: TCTTCAGAGCCAGGGAAGCCACCACCAAT	60	8	8

*1, Kim *et al*. 1999; 2, Horlein *et al*. (1993); 3, S Daval *et al*. unpublished; 4, Takao *et al*. (1988); 5, Burt & Paton (1991); 6, Caspers *et al*. (1997); 7, Lemonnier *et al*. (1991); 8, Maguire *et al*. (1986); 9, Stone *et al*. (1985); 10, Peter *et al*. (1989); 11, Wistow *et al*. (1988); 12, Schliess & Stoffel (1991); 13, Clausse *et al*. (1996); 14, Johnson & Bulfield (1992).

†Primers amplify exonic sequence only.

‡Primers to amplify intron sequence in *Brachyramphus marmoratus* designed from original sequences in Friesen *et al*. (1997).

§Primers to amplify intron sequence in *B. marmoratus* designed from original sequences in Friesen *et al*. (1999).

Appendix 1 *Continued*

(b)	Source species					Chromosomes screened	
Locus	Accession No.	Species	Ref.*	Primer sequences (5′–3′)	Annealing temp. (°C)	Pied	Collared
BMC4	AF005377	*Manorina melanophrys*	1	F: GATAGGAGACTGAGAGACTGTCCC R: TTTCTGAAGGGTTAGCTACAGACC	55–45†	6	8
HrU3	X84088	*Hirundo rustica*	2	F: CACTGGCTCTAGGCTGTCATC R: CTGTCCCATGTCAGGCCAGTC	59	12	8
Mcyμ5	U82389	*Malurus cyaneus*	3	F: GAGACTTTGTGTTGCTGTTAGG R: TTTGCATAGTAAGAATGAGAACAC	55–45†	8	6
Pocc8	U59119	*Phylloscopus occipitalis*	4	F: GCATGTCTCTTCAGACATCTGC R: ATGTAGAGCTCCCATGGTGG	55–45†	8	8
ZL35	AF076670	*Zosterops lateralis*	5	F: CCCAGGCATTTGCTGTAACTCG R: GCTGGTGCTGTCAGTCCCAC	57	10	10
ZL38	AF076672	*Z. lateralis*	5	F: CCTCAAGGTTAACCACTATAGAC R: GTAGTAGTATCTTCTGCATCAAGG	58	4	4

*1, Painter *et al*. (1997); 2, Primmer *et al*. (1995); 3, Double *et al*. (1997); 4, Bensch *et al*. (1997); 5, Degnan *et al*. (1999).
†Touchdown PCR protocol. See Koskinen & Primmer (1999).

(c)	Source species					Chromosomes screened	
Locus	Accession No.	Species	Ref.*	Primer sequences (5′–3′)	Annealing temp. (°C)	Pied	Collared
Fa45B	AJ299593	*Ficedula albicollis*	1	F: CAGGGAGCATTTTTAAACAGTG R: GGCAAAGTAGCACTCAGGGT	57	2†	12
Fh33	AJ299605	*F. hypoleuca*	1	F: GCAGCCTGTGATATAATTTCC R: GTGCCTTGACAACAACTTCTT	57	2†	2
Fh33B	AJ299615	*F. hypoleuca*	1	F: GTAGCATTCCAGAGTAGCACAGTC R: CCACTTCAAGTTTCTCCCTGAT	57	8	4
Fh45	AJ299632	*F. hypoleuca*	1	F: GCTTACTTCTGTGTTGTTATTGTAA R: GAGTTTCAGTGTCTTTCATTCAT	58	6	6
Fh49	XXxxxxxx	*F. hypoleuca*	2	F: GGAAGGATTTTTGAAGTCTGGT R: GCCTGACTGTGACATGCTTG	57	2†	2
Fh61	AJ299646	*F. hypoleuca*	1	F: GGAACTCCAGCTTTGAGTGTA R: CCATTTTCTTGACGATTTAGG	57	4	10
Fh63	AJ299655	*F. hypoleuca*	1	F: TGTGAGGGTCCATAGCATCC R: GCCTCATAGAACCAAGGAGACA	57	8	6
Fh69	AJ299668	*F. hypoleuca*	1	F: GCCAGCAGTGTTCTCTC R: CTGCAAGCTCAACACAA	58	6	4

*1, Sætre *et al*. 2001; 2, this study.
†An additional 12–14 individuals of each species were screened for polymorphism using SSCP (see Sætre *et al*. (2001) for methodological details), but no polymorphism was revealed.

