Hitchhiking and recombination in birds: evidence from *Mhc*-linked and unlinked loci in Red-winged Blackbirds (*Agelaius phoeniceus*)

SCOTT V. EDWARDS1,2* AND MEGAN DILLON1†

1 Department of Biology, University of Washington, Seattle, WA 98195, USA
2 Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138 USA

(Received 20 February 2004 and in revised form 8 September 2004)

Summary

Hitchhiking phenomena and genetic recombination have important consequences for a variety of fields for which birds are model species, yet we know virtually nothing about naturally occurring rates of recombination or the extent of linkage disequilibrium in birds. We took advantage of a previously sequenced cosmid clone from Red-winged Blackbirds (*Agelaius phoeniceus*) bearing a highly polymorphic *Mhc* class II gene, *Agph-DAB1*, to measure the extent of linkage disequilibrium across ~40 kb of genomic DNA and to determine whether non-coding nucleotide diversity was elevated as a result of physical proximity to a target of balancing selection. Application of coalescent theory predicts that the hitchhiking effect is enhanced by the larger effective population size of blackbirds compared with humans, despite the presumably higher rates of recombination in birds. We surveyed sequence polymorphism at three *Mhc*-linked loci occurring 1.5–40 kb away from *Agph-DAB1* and found that nucleotide diversity was indistinguishable from that found at three presumably unlinked, non-coding introns (*β*-actin intron 2, *β*-fibrinogen intron 7 and rhodopsin intron 2). Linkage disequilibrium as measured by Lewontin’s *D'* was found only across a few hundred base pairs within any given locus, and was not detectable among any *Mhc*-linked loci.

Estimated rates of the per site recombination rate *r* derived from three different analytical methods suggest that the amounts of recombination in blackbirds are up to two orders of magnitude higher than in humans, a discrepancy that cannot be explained entirely by the higher effective population size of blackbirds relative to humans. In addition, the ratio of the number of estimated recombination events per mutation frequently exceeds 1, as in *Drosophila*, again much higher than estimates in humans. Although the confidence limits of the blackbird estimates themselves span an order of magnitude, these data suggest that in blackbirds the hitchhiking effect for this region is negligible and may imply that the per site per individual recombination rate is high, resembling those of *Drosophila* more than those of humans.

1. Introduction

The increase in genome-scale data has made many previously intractable population genetic problems become more amenable to study using DNA sequence data. One such problem which was articulated decades ago (Felsenstein, 1974; Maynard-Smith & Haigh, 1974; Thomson, 1977) but which has only recently been subjected to rigorous quantitative analysis is the phenomenon of genetic hitchhiking. Hitchhiking arises when the evolutionary trajectory of selectively neutral regions in the genome is influenced by the trajectories of closely linked loci that are the actual targets of selection. Although many factors are involved, the degree of hitchhiking largely depends upon the recombination rate and the physical distance between a neutral locus and its linked selected gene, where recombination increasingly decouples the evolutionary histories of the two loci, causing the effects...
of selection on the target locus to be minimized at the neutral locus (Barton, 2000). Thus, estimating the extent of recombination in DNA sequences can contribute to predictions of the extent of genome-or chromosome-wide hitchhiking.

The resurgence of interest in hitchhiking is partly due to the discovery of a correlation between genomic diversity and local recombination rate in Drosophila and humans (Aquadro et al., 1994; Nachman, 2001, 2002; Payseur & Nachman, 2002b), a correlation which is presumably modulated by hitchhiking of neutral regions on either positively or negatively selected loci (but see Lercher & Hurst, 2002; Hellmann et al., 2003). Because it has the potential to amplify the effects of selection at a single site across large regions of chromosomes, hitchhiking and recombination can play a major role in influencing naturally occurring levels of genetic variation, haplotype structure and genome evolution (Posada et al., 2002). Indeed, another means of evaluating the extent of hitchhiking is by examining patterns of linkage disequilibrium (LD) between sites thought to be under selection and adjacent regions. Recently, discrepancies between levels of LD and estimated recombination rates have suggested that additional factors, such as population structure and selection, could also play important roles in modulating LD and haplotype structure (Pritchard & Przeworski, 2001; Stumpf & McVean, 2003). What is even less clear is how prevalent hitchhiking is in species that are not models for genetic analysis.

Few areas of the genome provide conditions more conducive to hitchhiking as a mode of evolution than the major histocompatibility complex (Mhc), a large multigene family encoding proteins that play an integral role in the adaptive immune response (Klein et al., 1993; Edwards & Hedrick, 1998; Yeager & Hughes, 1999). The high degree of polymorphism found in Mhc genes has been attributed to balancing selection, in the form of either heterozygote advantage or frequency-dependent selection (Garrigan & Hedrick, 2003). This selection is thought to be mediated by the rapidly evolving parasites that are bound by Mhc gene products, and in recent years several convincing examples of Mhc–parasite associations in humans and other vertebrates have been documented (reviewed in Hill, 1998; Dean et al., 2002). The human MHC (HLA) has long been a focus of studies of linkage disequilibrium, and several studies have found that the LD detected is probably maintained not so much by lack of recombination in the region but by selection maintaining distinct multilocus haplotypes (Begovich et al., 1992, 2001; Hollenbach et al., 2001). Regions of the genome surrounding Mhc genes have been characterized in humans as having several megabase-long tracts of DNA in linkage disequilibrium (Jeffreys et al., 2001; Walsh et al., 2003), as well as discrete hotspots of recombination (Crawford et al., 2004; Ptak et al., 2004). By contrast, analysis of non-coding regions in the vicinity of human and mouse Mhc genes shows that the effect of hitchhiking can be substantial, sometimes raising the level of nucleotide diversity 10-fold over the ‘background’ level found at neutral regions unlinked to the Mhc (Meager & Potts, 1997; Satta et al., 1999).

For a variety of structural and evolutionary reasons, the Mhc of birds, in contrast to human Mhc, may conform to an alternate evolutionary model. At less than 100 kb, the Mhc (B-complex) in chickens is much smaller and more gene-dense than that in mammals (Kaufman, 1999; Kaufman et al., 1999). This fact alone might indicate that the Mhc of birds would be much more subject to hitchhiking phenomena, both because of its small size and because of the high density of potential targets of selection, in the form of coding regions (McQueen et al., 1998; Payseur & Nachman 2002a). In addition, the chicken Mhc is located on one of the microchromosomes, which are now known to be about 3–12 Mb in length (reviewed in Burt, 2002). This characteristic could potentially increase the extent of hitchhiking for the above reasons, but could also decrease hitchhiking because of the obligatory crossing over that takes place on all avian microchromosomes every meiosis. Consistent with this idea is the finding that physical and genetic linkage data in chickens places 1 cM between on average 350 and 440 kb (Smith & Burt, 1998), whereas the average value for humans is 1 Mb per 1 cM.

In chickens, the Mhc appears to undergo very little recombination as evidenced by the stability of multilocus haplotypes in surveys of thousands of meioses in pedigrees (Hala et al., 1988). Indeed, levels of nucleotide variation at some Mhc-linked loci appear high, although there has been little quantification of levels of diversity or comparison with regions unlinked to the Mhc in chickens (Kaufman, 1995, 1999; Kaufman & Salomonsen, 1997). By contrast, analysis of sequence diversity in a chicken class I locus revealed small-scale evidence for recombination, although no attempt was made to estimate rates (Hunt et al., 1993). Thus existing evidence is contradictory and it is unclear whether the higher recombination rates in birds or the avian Mhc may ultimately render hitchhiking a weak force in moulding genetic diversity. Although surveys of haplotype stability in chicken pedigrees are a powerful method for detecting recombination, an alternative approach is the analysis of DNA sequence data in outbred populations, which has the advantage of averaging over many more thousands of meioses in the history of a species, even with relatively small sample sizes. Despite the completion of the chicken genome project, significantly
we know of no study that has systematically investigated the extent of genetic hitchhiking in chickens or any other bird, and the applicability of findings from chickens to other avian species is not known.

Among the passerine birds (Passeriformes), the Mhc genes of Red-winged Blackbird (Agelaius phoeniceus) are particularly well studied. Red-winged Blackbirds is an ecologically well studied species belonging to an evolutionarily well studied New World clade (Freeman & Zink, 1995; Beletsky, 1996). We have cloned three full-length Mhc class II genes from this species, only one of which, Agph-DAB1, exhibits the high polymorphism and evidence for selection expected of functional Mhc genes (Edwards et al., 1998). A first look at polymorphism in the intron adjacent to the Agph-DAB1 peptide-binding region (PBR) revealed that the level of intron diversity was considerably lower than that found in the PBR, although this level was still substantial (Garrigan & Edwards, 1999). Paradoxically, this intron exhibited a signature (a significantly negative value of Tajima's D statistic (Tajima, 1989), of directional selection, or possibly of population growth, rather than the strongly positive Tajima's D expected of a locus linked to one under balancing selection. Although estimates of recombination rates in this gene were high, the absence of estimates of both neutral levels of nucleotide diversity made it difficult to determine the evolutionary forces creating the discordance in polymorphism between the two loci. Additionally, the situation upstream of the Agph-DAB1 PBR could in principle be dramatically different from that exhibited by the downstream intron that was investigated.

The sequencing and detailed characterization of the cosmid (RWCos3) on which Agph-DAB1 is found (Gasper et al., 2001) paves the way for extending the study of hitchhiking in songbirds. The present study aims to quantify the importance of hitchhiking in blackbirds by measuring diversity as well as recombination rates throughout the 45 kb region upstream of Agph-DAB1 and by comparing these to levels of diversity and recombination found in three additional nuclear introns, presumably unlinked to the blackbird Mhc. Because of their large effective population sizes and lack of significant genetic structure (Ball et al., 1988; Garrigan & Edwards, 1999), Red-winged Blackbird is an appropriate species in which to relate evolutionary forces such as recombination specifically to LD and hitchhiking; in addition to being influenced by recombination, LD is known to be caused by bottlenecks, population structure and other demographic forces, and so studying LD in a species with little structure can help reduce the possible explanations for the patterns of disequilibrium found (Slatkin & Wiehe, 1998; Reich et al., 2002; Wakeley & Lessard, 2003). Finally, our data add to a small but growing data base of measurements of non-coding single-nucleotide polymorphism (SNP) diversity in birds (Congdon et al., 2000; Brumfield et al., 2003).

2. Methods

(i) Quantitative model of hitchhiking

We used equations 1–7 of Takahata & Satta (1998) to explore quantitative predictions for the extent of hitchhiking in a typical avian species. These equations permit a prediction of the nucleotide diversity at a neutral locus (\(\pi = 2 \mu T\), where \(\mu\) is the neutral mutation rate per site per generation and \(T\) is the coalescence time in generations) separated from a target under strong symmetrical balancing selection (e.g. an Mhc PBR) of strength \(s\) by a rate of recombination \(c\) scaled by the sample size. The units of \(c\) are recombination events per site per generation per individual. For these analyses we relied heavily on estimates for population genetic parameters for Red-winged Blackbird (Garrigan & Edwards, 1999) and chicken (Smith & Burt, 1998). Briefly, we calculated \(T\) as:

\[
T = \frac{1}{n} T_w + \left(1 - \frac{1}{n}\right) T_b, \tag{1}
\]

where

\[
T_w = \frac{2 N_s (a + c^*n)}{(n + 2 N_s a (a + c^*))} \tag{2}
\]

and

\[
T_b = T_w + \frac{n - 1}{2(a + c^*)} \tag{3}
\]

and \(n\) is the sample size of alleles, \(\mu\) is the neutral mutation rate, and \(T_w\) and \(T_b\) are coalescence times within and between ‘allelic lineages’. These coalescence times within and between lineages were originally parameterized for the human HLA, but are used here for simplicity. The rate of loss of allelic lineages, and hence of nucleotide diversity, is captured in the parameter \(a\), which summarizes the relative strength of balancing selection versus genetic drift due to finite \(N_e\), or effective population size (see Takahata & Satta 1998) equation 1). \(\mu\) was centred around \(10^{-8}\), a value that receives some support from avian studies (Axelsson et al., 2004), but sensitivity to this parameter was gauged by varying it from \(10^{-8}\) to \(10^{-10}\) (see legend to Fig. 2). We generated a range of curves corresponding to values for \(N_e\), from humans (10 000) and blackbirds (\(\sim\) 50 000–10 0000 based on previous mitochondrial DNA studies: Ball et al., 1988). The selection intensity \((s)\) was varied such that
the product 2Ns for the human curves in Fig. 2 ranged between 200 and 20,000, values that encompass the range of values receiving support from recent analyses of HLA data (Satta et al., 1994; Slatkin & Muirhead, 2000). The same range of values for s (0.01–0.1) was applied to the avian curves in Fig. 2, as well as s = 0.001 to explore the consequences of weaker selection in birds. We applied an average physical length per centimorgan of 350 kb in chickens (Smith & Burt, 1998) to the blackbird data set. While this assumption is surely naïve, it is a good starting point.

(ii) DNA samples
Blackbird tissue samples were obtained from the Genetic Resources Collection of the Burke Museum, University of Washington, Seattle, with the exception of two samples from Florida from Garrigan & Edwards’s (1999) study. Following a proteinase K tissue digestion, DNA was extracted using a standard phenol-TE dialysis method (Sambrook & Russell, 2001). Although sample sizes varied depending on the locus being studied (Table 1), for most loci in this study 18 chromosomes (9 individuals) were sampled. To encompass as much of the geographic range of this species as possible, we used birds from throughout North America (Washington state, Colorado, Oregon, Florida, Kentucky and Nicaragua; Appendix).

(iii) Genetic regions and molecular protocols
Sequencing of the cosmids containing Agph-DAB1 (RWCos3) revealed that the gene is located on one end of the clone and allowed us to characterize in detail the genomic region on the side of the gene opposite that of intron 2 (Fig. 1; Gasper et al., 2001). The sequence of this cosmid spans 45 kb and encodes an additional Mhc class II B gene, Agph-DAB3, as well as three other likely functional genes. RWCos3 was also found to have large duplicated regions paralogous to regions on a second cosmid, RWCos10 (Edwards et al., 2000). The new loci chosen for this study were confined to those regions known not to be duplicated or paralogous. The first region consisted of the previously published Mhc class II exon and intron data from the Agph-DAB1 locus (Garrigan & Edwards, 1999; accession numbers

Table 1. Sequence statistics for seven loci in Red-winged Blackbird

<table>
<thead>
<tr>
<th>Mhc-linked loci</th>
<th>Unlinked to Mhc</th>
<th>β-fibrinogen (BFIB7)</th>
<th>β actin (BACT2)</th>
<th>Rhodopsin intron 2 (RDP2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agph-DAB1 exon 2</td>
<td>Agph-DAB1 intron 2</td>
<td>Agph-DAB1 promoter</td>
<td>Agph-DAB3</td>
<td>Rhodopsin intron 2</td>
</tr>
<tr>
<td>No. of chromosomes</td>
<td>Base pairs sequenced</td>
<td>Base pairs sequenced</td>
<td>Base composition</td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>89</td>
<td>23:1:30:4</td>
<td>23:1:30:4</td>
<td>23:1:30:4</td>
</tr>
<tr>
<td>33</td>
<td>245</td>
<td>16:7:45:0</td>
<td>16:7:45:0</td>
<td>16:7:45:0</td>
</tr>
<tr>
<td>35:8:23:9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25:1:33:9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A : C</td>
<td>G : T</td>
<td>Ti/Tv, α</td>
<td>Var. sites (S)</td>
<td>No. of haplotypes</td>
</tr>
<tr>
<td>23:1:30:4</td>
<td>30:7:15:9†</td>
<td>1, 0-31</td>
<td>22</td>
<td>8</td>
</tr>
<tr>
<td>16:7:45:0</td>
<td>30:7:20:4</td>
<td>1, 0-17</td>
<td>36</td>
<td>23</td>
</tr>
<tr>
<td>13:1:35:9</td>
<td>31:5:17:9</td>
<td>1, ∞</td>
<td>3</td>
<td>23</td>
</tr>
<tr>
<td>23:1:35:9</td>
<td>31:5:17:9</td>
<td>1, ∞</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>16:7:26:2</td>
<td>31:5:17:9</td>
<td>1, ∞</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>25:1:33:9</td>
<td>31:5:17:9</td>
<td>1, ∞</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>Base composition:</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>A : C</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>G : T</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Ti/Tv, α</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Var. sites (S)</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>No. of haplotypes</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Watterson’s θ</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>0:071</td>
<td>0:040</td>
<td>0:00254</td>
<td>0:00254</td>
<td>0:01240</td>
</tr>
<tr>
<td>Nuclear diversity (π)</td>
<td></td>
<td></td>
<td></td>
<td>0:00614</td>
</tr>
<tr>
<td>1:010</td>
<td>0:018</td>
<td>0:00217</td>
<td>0:0022</td>
<td>0:00775</td>
</tr>
<tr>
<td>Tajima’s D</td>
<td></td>
<td></td>
<td></td>
<td>0:00245</td>
</tr>
<tr>
<td>1:9*</td>
<td>−1:7*</td>
<td>−0:40407</td>
<td>−0:73968</td>
<td>−0:22124</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1:08303</td>
<td>−0:84182</td>
<td>−1:50776</td>
</tr>
</tbody>
</table>
| Ti/Tv and α indicate the transition/transversion ratio and among-site rate heterogeneity for the locus as determined in MODEL TEST.† Not significantly different from equal base frequencies as determined by likelihood ratio tests in MODEL TEST; *P < 0.05. a As determined by the computer program PHASE.

Fig. 1. Schematic diagram of cosmids RWCos3 (Gasper et al., 2001). Black bars above genes indicate Mhc-linked regions amplified.
AF105404–AF105420 and AF133292–AF133297). The next three regions are defined as Mhc-linked and occur on the cosmid RWCos3 along with the Agph-DAB1 locus (Fig. 1). These included an anonymous, apparently non-coding region 40 kb upstream of Agph-DAB1 (‘40 kb locus’; a BLAST search of this locus revealed no significant hits except the cosmid from which it came); exon 2 of Agph-DAB3 (‘DAB3’), which was also surveyed for polymorphism in Gasper et al. (2001), but for a different set of individuals; and the promoter region of Agph-DAB1 (‘DAB1 promoter’). The primers 3RW-1035F (5’-TACTGGTAGCAACAGCCATT-3’) and 3RW-1934R (5’-AGCAGCTAATAGCCATGCTT-3’) amplified the 40 kb locus; primers in Gasper et al. (2001) were used to amplify DAB3, and primers DAB1-RR-F (5’-GCGTGGATGATCCCAGTGAC-3’) and DAB1-RR-R (5’-TCTCATTGGCTGCAACGCCTC-3’) amplified the DAB1 promoter. Additionally, three non-coding loci, presumably unlinked to the blackbird Mhc, were amplified: intron 2 of rhodopsin (‘RDP2’), intron 7 of the β-fibrinogen gene (‘BFB7’) and intron 2 of the β-actin gene (‘BACT2’), all of which have been used previously for avian evolutionary studies and primes for which are available (Prychitko & Moore, 1997; Waltari & Edwards, 2002).

All regions were directly sequenced on an ABI 377 automated sequencer with the above amplification primers with the exception of the 40 kb locus, which was directly sequenced with the primers 3RW-1188F (5’-TCTCAACAGAGGTGAGATAA-3’) and 3RW-1840R (5’-ACCTCCTGTGGTCCCTAGGTT-3’). In order to conduct HKA tests of neutrality, which require an outgroup species (Hudson et al., 1987), we attempted to amplify these loci from DNA from Brown Cowbird (Molothrus ater) and Bullock’s Oriole (Icterus bullockii). Because only the RDP2 and BFB7 loci amplified and sequenced reliably from these species, and because the HKA tests conducted did not reject neutrality, we do not report data from these tests further. All new sequences (those unlinked to the Mhc) have been submitted to the GenBank data base under accession numbers AY714388 (RDP2), AY714390 (BF1B7), AY714391 (BACT2). SNPs in these and the Mhc-linked loci have been submitted to the NIH dbSNP data base under SNP accession numbers 28514110–28514117 (BACT2) and 28514123–28514124 (RDP2), 28514125–28514127 (DAB1 promoter), 28514128–28514129 (DAB3), 28514130–28514147 (BF1B7), 28514148–28514161 (40 kb).

(iv) Measures of polymorphism and neutrality
As an aid to manual sequence inspection in the search for polymorphic SNPs, the raw chromatograms were analysed using Polyphred (Nickerson et al., 1997; Rieder et al., 1998), which uses details of chromatogram peak height, morphology and quality to identify heterozygous sites in diploid PCR products (see Brumfield et al., 2003). Genotypes of heterozygous SNPs were then resolved into haplotypes using PHASE v. 2.0, a program that yields Bayesian estimates of haplotypes and their frequencies from genotypic data under the assumption of random mating (Li & Stephens, 2003; Stephens & Donnelly, 2003). Phases of the exon 2/intron 2 data of Garrigan & Edwards (1999) had already been resolved via cloning of PCR products, in the case of the three additional Mhc-linked loci analysed here, we incorporated the phase information from one individual (individual 7, Appendix) provided by the original cosmid sequence into our PHASE analysis. Haplotypes resolved using PHASE generally had high probabilities associated with heterozygous sites and were robust to assumptions of recombination or lack thereof (Appendix). The program DNAsp (Rozas & Rosas, 1997) was used to calculate basic sequence statistics, such as nucleotide diversity (θ) and Tajima’s D. Although no phylogenetic analyses were done, we nonetheless used PAUP*4.0b10 (Swofford, 1999) and the program MODELTEST (Posada & Crandall, 2001) to collect basic evolutionary information on these predominantly non-coding loci, such as base composition, transition – transversion ratio and rate heterogeneity among sites (Table 1).

(v) Linkage disequilibrium and recombination
Haplotypes generated using PHASE were used to calculate D′, a common measure of linkage disequilibrium (Lewontin, 1995), between polymorphic sites of the Mhc linked loci using the program LD plot, provided by S. Schaeffer (see Schaeffer et al., 2001). Because the set of individuals used by Garrigan & Edwards (1999) and in this study were only partially overlapping, we only estimated LD for sites within each of these two data sets. Because of our small sample size for some loci, we examined statistical significance with and without Bonferroni correction (Rice, 1989).

As recommended (Posada & Crandall, 2001), we used multiple computer programs to estimate the population recombination parameter, ρ=4Nc from the collected sequence data (see recent reviews see Posada, 2002; Rokas et al., 2003). The program Ldhat (Hudson, 2001; McVeay et al., 2002) is a maximum likelihood method that calculates the coalescent likelihood of every pair of polymorphic sites, and generates a composite likelihood function of ρ per locus across all sites. ρ per site was calculated by dividing the Ldhat ML estimate by the number of sites in the locus. We also used PHASE to estimate ρ per site while simultaneously estimating haplotypes.
(Stephens et al., 2001b; Li & Stephens, 2003; Stephens & Donnelly, 2003). PHASE generates a posterior probability distribution of ρ and can be checked for convergence. When estimating ρ we ran PHASE with varying random number seeds and starting points for ρ, usually 10,000. Both PHASE and Ldhat analyse only single sets of linked sites at a time. Finally, we used the Markov Chain Monte Carlo method RECOMBINE (Kuhner et al., 2000), as implemented in the program LAMARC (http://evolution.genetics.washington.edu/lamarc.html), to estimate ρ for individual loci as well as to generate an overall value of ρ across all loci analysed, as well as associated values of θ = 4 Neμ. Again, multiple runs and increases in Markov chain length and number were used to ensure convergence of parameter values. Although all of these programs assume neutrality in their estimation of ρ from sequence data, we nonetheless analysed some loci clearly violating neutrality (e.g. Agph-DAB1) for comparison with loci determined to be evolving neutrally. Two types of multilocus analyses were done. First, the three Mhc-linked loci spanning ~40 kb were analysed simultaneously using PHASE and Ldhat; in the latter procedure, the average value of θ across the three loci was applied to the entire ~40 kb region. Second, a multilocus average value for ρ was estimated using LAMARC. Here, as many loci as possible were included, regardless of level of polymorphism, so as to maximize precision of the global estimate.

3. Results

(i) Theoretical expectations

Application of the hitchhiking model of Takahata & Satta (1998) and recombination parameters from chickens (Smith & Burt, 1998) suggest that the hitchhiking effect should extend along the chromosome to similar degrees in birds and humans. In our exploration of parameter space, we varied only Ne and s. Like Takahata & Satta (1998), we found that the product Ne s was largely responsible for setting the leftmost ‘height’ of the curve and the level of neutral diversity close to the target of selection (on the left-hand end of Fig. 2), whereas Ne alone influenced the level of diversity far from the target of selection (on the right-hand end of Fig. 2). The values corresponding to human populations (lowest curve) fell to neutral levels within about 6 kb, whereas, depending on the values of Ne and s assumed for the avian curves, nucleotide diversity could remain higher than the neutral expectation at least this far from the target of selection. Ne seemed to influence the rate of decline to neutral levels more than did the recombination rate.

(ii) Polymorphism and haplotyping

Of the 1135 nucleotides resequenced across the three Mhc-linked loci (DAB1 promoter, DAB3, 40 kb
locus), 19 sites were polymorphic, yielding an overall $\theta_w$ (Watterson’s $\theta$) of 0.006, a value that is substantially higher than that found in humans but similar to levels in other birds (Wang et al., 2003; see Table 1). Across 990 sites resequenced for the three introns unlinked to the $\text{Mhc}$ ($\text{BFIB7}$, $\text{BACT2}$, $\text{RDP2}$), there were a total of 28 polymorphic sites, yielding an overall $\theta_w$ of 0.007, which is indistinguishable from the average value for the $\text{Mhc}$-linked loci. $\theta_w$ for both $\text{Mhc}$-linked and presumably unlinked loci were substantially lower than those for the $\text{Agph-DAB1}$ exon 2 and intron 2 (0.101, 0.018, respectively). Nucleotide diversity ($\pi$) was similar to $\theta_w$ for all loci, except for the $\text{Agph-DAB1}$ intron, which showed a smaller $\pi$ value given $\theta_w$ and therefore rejected neutrality (Garrigan & Edwards, 1999). Overall $P$ values were consistent with theoretical expectations resulting from parameter combinations of high $N_e$ and/or high $s$ (Fig. 2). For the two protein-coding loci synonymous diversity ($\pi_s$) was similar to non-synonymous diversity ($\pi_N$), although both these values for $\text{Agph-DAB1}$ were small ($\pi_s$, 0.00172; $\pi_N$, 0.00146; $\text{Agph-DAB1}$). None of the other six loci rejected the hypothesis of neutrality using Tajima’s $D$ test and only one locus (40 kb) was positive for Tajima’s $D$ (Table 1). The $D$ value for RDP2, although nonsignificant, was substantial and may warrant further investigation. The interspecific divergence for these loci largely predicted the within-blackbird level of polymorphism, and none was able to reject neutrality by the HKA test.

The genotypic variation was resolved into haplotypes using PHASE, with and without the assumption of recombination. The variation in the three $\text{Mhc}$-linked loci was resolved simultaneously into 17 distinct haplotypes out of a total 18 possible, indicating high haplotypic diversity (Appendix). Although the most probable multilocus haplotypes were the same whether or not information from the two known haplotypes was included, probabilities of individual sites within haplotypes generally increased due to inclusion of these known haplotypes (not shown). The $\beta$-fibrinogen intron ($\text{BFIB7}$) also had a high haplotypic diversity, with 13 out of a possible 18 haplotypes recovered, each of which was found in only 1 or 2 copies. For the two additional loci unlinked to $\text{Mhc}$ genes, there was a single common inferred haplotype and a smaller number of total haplotypes: $\text{BACT2}$ had 9 inferred haplotypes, the most common of which was inferred to have occurred 12 times, and $\text{RDP2}$ had 3 haplotypes, with the most common found 16 times.

(iii) Linkage disequilibrium

Analysis of the $\text{Agph-DAB1}$ exon/intron data set of Garrigan & Edwards (1999) suggested high levels of LD within the exon, but fewer sites in the intron in LD with other sites (Fig. 3, left). Of the 50 sites exhibiting some level of LD with other sites in the region ($P<0.05$), 20 (40%) were in the exon, even though only 89 of the 337 total sites (26%) were exonic. High levels of LD in the exon were probably generated by a gene conversion event, discussed by Garrigan & Edwards (1999), that caused a suite of sites (3, 4, 7, 28, 31, 35, 62–64) to covary highly non-randomly. There
was nonetheless a cluster of sites in the intron (139, 147, 148, 182, 194) that showed substantial LD with each other and with several exonic sites (83, 84, 89) without Bonferroni correction, suggesting that, despite an inference of approximately 8 recombination events in the region by Garrigan & Edwards (1999), LD might extend throughout the region. Still, with Bonferroni correction, the region experiencing significant LD extends only to site 65 in the exon, and does not include any intronic sites.

LD was measured across the three additional loci spanning ~40 kb of the blackbird cosmid. Significant levels of LD, however, were found primarily within the first locus, 40 kb, extending in patches throughout all 14 variable sites (Fig. 3, right) spanning 419 bp ($P < 0.05$), or across the first 5 variable sites, spanning only 96 bp with Bonferroni correction. $P < 0.05$ was suggested between four sites in 40 kb (155, 263, 321, 330) and site 40868 in the DAB1-promoter, but these associations vanished after Bonferroni correction.

(iv) Estimates of recombination rate ($\rho$)

The program Ldhat was used to analyse five data sets. The RDP2 and DABI-promoter loci were not analysed individually because of the low level of polymorphism (2 and 3 segregating sites, respectively). Maximum likelihood estimates of $\rho$ per site varied from a low of 0.045 for Agph-DAB1 to 0.225 for the BFIB7 locus (Table 2). However, the likelihood curve for BFIB7 was very flat, indicating little confidence in the upper bound of the estimate (Fig. 4d). By contrast, the likelihood curve for Agph-DAB1 was much tighter (Fig. 4a), with most of the signal coming from variation in the exon rather than from the intron. The number of segregating sites within a locus only partly explained the variation in likelihood curves, since BACT2, which had the fewest variable sites, had a detectably sharper curve than did BFIB7 (not shown).

A multilocus analysis of the three Mhc-linked loci (40 kb, DAB1-promoter, DABS) for the LAMARC analyses, ‘Multilocus’ refers to analysis of all six neutral loci (all except Agph-DAB1 exon and intron), some of which are unlinked to Agph-DAB1. For the PHASE and LAMARC analyses, the 95% CIs are those recovered from a single representative run.

Table 2. Estimates of recombination rate ($\rho$) in Mhc-linked and unlinked loci in Red-winged Blackbird

<table>
<thead>
<tr>
<th></th>
<th>Mhc-linked</th>
<th>Unlinked</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Agph-DAB1</td>
<td>BFIB7</td>
</tr>
<tr>
<td></td>
<td>exon 2/intron 2</td>
<td>intron 2</td>
</tr>
<tr>
<td></td>
<td>40 kb locus</td>
<td>(BFI7)</td>
</tr>
<tr>
<td></td>
<td>Multilocusa</td>
<td>(BACT2)</td>
</tr>
<tr>
<td>No. of sites spanned</td>
<td>337</td>
<td>422</td>
</tr>
<tr>
<td>Ldhat (ML score)</td>
<td>–38 002.66</td>
<td>329</td>
</tr>
<tr>
<td>$\rho$ (per locus)</td>
<td>0.15-124</td>
<td>0.0059–0.0088b</td>
</tr>
<tr>
<td>$\rho$ (per site)</td>
<td>0.045</td>
<td>0.0076–0.0512b</td>
</tr>
<tr>
<td>CI (per site)c</td>
<td>0.0412–0.056</td>
<td>0.0048–0.0586b</td>
</tr>
<tr>
<td>$\rho/\theta$</td>
<td>0.96</td>
<td>0.009–0.022c</td>
</tr>
<tr>
<td>PHASE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\rho$ (per site)</td>
<td>0.0016–0.0241</td>
<td>0.0016–0.020b</td>
</tr>
<tr>
<td>95% CI</td>
<td>6.604</td>
<td>0.002–0.022c</td>
</tr>
<tr>
<td>$\rho/\theta$</td>
<td>0.127</td>
<td>0.0005–0.0254</td>
</tr>
<tr>
<td>LAMARC (ML)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\theta$ 95% CI</td>
<td>n.d.</td>
<td>0.008–0.022b</td>
</tr>
<tr>
<td>$\theta/\theta$ 95% CI</td>
<td>n.d.</td>
<td>0.008–0.022b</td>
</tr>
<tr>
<td>$\rho/\theta$ 95% CI</td>
<td>n.d.</td>
<td>0.009–0.022b</td>
</tr>
<tr>
<td>$\rho/\theta$ per site</td>
<td>n.d.</td>
<td>0.0005–0.0254</td>
</tr>
</tbody>
</table>

n.d., not done.

For the Ldhat and PHASE analyses, the ‘Multilocus’ category refers to analysis of the Mhc-linked loci (40 kb, DABI-promoter, DABS). For the LAMARC analyses, ‘Multilocus’ refers to analysis of all six neutral loci (all except Agph-DAB1 exon and intron), some of which are unlinked to Agph-DAB1. For the PHASE and LAMARC analyses, the 95% CIs are those recovered from a single representative run.

Range of values indicates the range of ML estimates across multiple runs. CIs from Ldhat based on ML ± 2 likelihood units.
Table 2); as with Ldhat, the estimate for *Agph-DAB1* was the lowest among the loci (Fig. 4a, b), although with PHASE the estimate for this locus was an order of magnitude lower than for Ldhat. Whereas the smoothness of the likelihood surfaces for *Agph-DAB1* and the 40 kb locus were reasonable (Fig. 5a–d), the surface for *BACT2* was less smooth (Fig. 5g, h). The surface for *BFIB7* was distinctly bimodal, with several cycles getting caught in optima of very high values for \( r \), and others converging on exceedingly low values (Fig. 5e, f). This probably indicates conflicting signals in the data, substantial homoplasies, or perhaps recombination rates that are very high and difficult to estimate. Even so, the lower bound of the 95% confidence limits for most loci excluded a commonly accepted average rate of recombination in humans of 0.0004 per base pair. As with Ldhat, all the estimates of \( \rho / \theta_w \) except that for *Agph-DAB1* exceeded 1 (Table 2).

The recombination module within the LAMARC package returns estimated values of \( \theta_L \) and \( c / \mu \) for each locus, as well as an average estimate across all loci under the default assumption that \( c \) for all loci is equal. For this reason the loci with few segregating sites such as *RDP2* and *DAB3*, for which estimating individual recombination rates would be very difficult, were included, since data from additional loci can only improve the global estimate. The results for analyses in which haplotypes were estimated simultaneously were similar to those in which haplotypes were assumed to be known. Estimates of \( \theta_L \) using LAMARC were slightly higher than Watterson’s \( \theta \) for most loci (Table 2). Estimates of \( c / \mu \) were substantially lower than those calculated for values from Ldhat or PHASE, ranging from \( \sim 0.1 \) to 2.5 for most loci, but as high as 5.9 for *BACT2* in one run. Values of \( c / \mu \) for *RDP2*, *DAB1* promoter and *DAB3* fell outside this range (<0.000006; data not shown), but, as expected, there were very few segregating sites to estimate recombination for these loci singly. Estimates of \( \rho \) that were back-calculated from these values were generally lower than values from the other two methods and ranged from 0.0005 to 0.07 for those loci with reasonable information content (Table 2). The lower bound of these estimates is broadly consistent with recent estimates from humans.

We used the range of loci analysed using LAMARC to determine whether there was a correlation between \( \rho / \theta_w \) and \( \rho \). We found a strong correlation using values generated by LAMARC but a much weaker pattern using estimates from other methods (not shown).

### 4. Discussion

The rate and amount of recombination and extent of hitchhiking both have important consequences for
a variety of fields (Pritchard & Przeworski, 2001), including those in which birds are important models, yet virtually nothing is known of naturally occurring levels of recombination in any bird other than chicken. For example, recombination can cause difficulties when estimating nuclear gene genealogies (Hare, 2001), but can allow non-coding portions of the genome – of increasing importance to avian population geneticists – to escape the effects of hitchhiking. In principle recombination rates also have important consequences for animal behaviour: for example, Mhc genes have been suggested to serve as kin-recognition or mate choice loci in outbred populations (Manning et al., 1992; Jacob et al., 2002; Zelano & Edwards

Fig. 5. Bayesian analysis of recombination using PHASE. Graphs in the left-hand column show traces of ρ values throughout the cycles of analysis. Graphs in the right-hand column show the posterior probability distribution of estimates of ρ for each data set. (a), (b) Agph-DAB1; (c), (d) 40 kb locus; (e), (f) BFIB7; (g), (h) BACT2.
Hitchhiking in the Mhc of blackbirds

2002), yet the number of independently segregating Mhc loci will dictate the level of matching complexity achievable with a tightly linked set of kin-recognition loci (Grafen, 1990). In this paper we have explored the extent of genetic hitchhiking and used several methods to estimate rates of recombination within and between loci that are linked and presumably unlinked to an Mhc gene under strong balancing selection. We find that estimated amounts of recombination are high and generally exclude a commonly cited level of the number of crossovers per nucleotide per generation of 0.0004 in humans. Other estimates in humans would place \( \rho = 0.0004 \) at the upper bound of recombination rates (Stumpf & McVean, 2003; McVean et al., 2004), thereby accentuating the difference between blackbird and human results even more. These results are consistent with high average recombination rates in chickens but are at odds with observed levels of recombination in the Mhc (B-complex) in chickens.

(i) Extent of hitchhiking in the vicinity of Agph-DAB1

Garrigan & Edwards (1999) found slightly elevated levels of variation in the intron immediately adjacent to the PBR of Agph-DAB1, and this study revealed levels consistent with neutrality in the Agph-DAB1 promoter a mere \( \sim 600 \) bp from the PBR. Thus the levels of neutral levels of diversity predicted by analytical theory remain high much farther from the target of selection than is observed in our blackbird data. Satta & Takahata (1999) found a similar discrepancy between expected and observed extent of hitchhiking in the human HLA. The analytical theory predicted that for avian population parameters, levels of diversity should decline to neutral levels within about \( \sim 7 \) kb away from the target of selection. Although the theory necessarily corresponds to an idealized population – in particular recombination rates are assumed to be constant thought the hypothetical Mhc region – we were surprised at the similarity in predictions between birds and humans (Fig. 2). It appears that the effective population size is very important in modulating the hitchhiking effect: if we assume a 10-fold higher \( N_e \) in blackbirds, even with the same selection coefficient, theory predicts that the asymptote for neutral variation is achieved at least as far from the target of selection as in humans, despite a 3-fold higher recombination rate assumed from the chicken data (Smith & Burt, 1998). Satta et al. (1999) collected data indicating that silent diversity can extend in excess of 100 kb from the nearest target of selection in the human HLA region, and Grimsley et al. (1998) found enhanced variation at an HLA pseudogene residing \( \sim 200 \) kb from HLAA, a target of strong balancing selection. It is likely that the close linkage of many HLA genes – some occur within 40 kb of one another – could be responsible for this maintenance of silent diversity so far from the presumed targets.

(ii) Rates of recombination at blackbird loci

Our sampling scheme for sequencing in this study made estimation of recombination rates challenging. Recombination is best detected across long DNA sequences that will record a number of crossover events detectable by the presence of all pairs of two-locus combinations (Posada & Crandall, 2001; McVean et al., 2002). Thus, the fact that our loci were at most a few hundred base pairs long, and in some cases encompassed coding regions or regions possibly under strong stabilizing selection (Agph-DAB1 promoter, Agph-DAB3), resulted in low polymorphism, making estimation of recombination difficult. On the other hand the values of \( \theta \), particularly for the non-coding loci, were generally high compared with human population studies. For nearly any method, estimating very high or very low recombination rates is difficult, because the data necessary to discriminate between multiple extreme values of \( \rho \) are usually lacking. Thus several of the loci exhibited very flat likelihood curves when analysed using Ldhat. Nonetheless, such analyses could reject very low values for \( \rho \) in regions of the space where the likelihood curve rose steeply (Fig. 4).

Our cosmid sequence and multilocus analyses in principle permitted estimation of \( \rho \) across tens of kilobases. Estimates of \( \rho \) from such analyses were surprisingly similar to those within individual loci, and both PHASE and Ldhat were able to accommodate such data easily. A recent study of recombination rates in humans (Clark et al., 2003) employed similar sampling scheme with a small number of SNPs (4–10) sampled across \( \sim 50 \) kb regions of the genome. While the human study benefits from complete knowledge of all non-polyorphic sites in such regions, most available recombination estimation packages (except LAMARC) do not take information from such invariant sites, such as base composition, into account when estimating \( \rho \). Additionally, our sampling scheme for the Mhc-linked loci, which involved complete resequencing across selected loci within a larger region, avoided some of the problems associated with ascertainment bias that require special statistical procedures to ameliorate (Wakeley et al., 2001; Nielsen & Signorovitch, 2003).

The seven loci we surveyed exhibited a range of recombination rates. Surprisingly, for those loci with moderate or high levels of \( \theta \), the lowest estimated rates came from the Agph-DAB1 Mhc locus. The frequency of recombination and/or gene conversion in mammalian Mhc class II loci, though controversial, is generally thought to be high (Gyllensten et al., 1990; Begovich et al., 1992; Bergström et al., 1998;
Hogstrand & Bohme, 1999a, b), although some analyses suggest otherwise (Nei & Hughes, 1992; Gu & Nei, 1999). A recent survey of a Mhc class II locus in Peromyscus estimated recombination rates substantially higher than \( \theta \) (Richman et al., 2003b). Ohta (2000) analysed data from the blackbird Agph-DAB1 locus as well as several mammalian Mhc class II genes and found that the standard identity excess, a statistic that quantifies the ‘patchwork’ pattern resulting from frequent gene conversion and with properties similar to Tajima’s \( D \), was demonstrably higher for mammalian than for avian PBR exons. Visual inspection of aligned avian class II sequences confirms that the bold patchwork pattern found in, for example, mouse and Peromyscus class II genes (She et al., 1990; Richman et al., 2003a), is not as evident at avian class II loci (Edwards et al., 1998). Richman et al. (2003b) recently suggested that such motifs persist in rodent class II loci in the face of high recombination because of selection for combinations of motifs. It may be that recombination and gene conversion rates are in fact lower at avian than in mammalian class II loci, resulting in less of a patchwork pattern.

Throughout our study we have focused on measuring recombination in terms of the composite parameter \( \rho = 4Nc \), which necessarily confounds \( N_e \) and \( c \). Thus our finding of high values of \( \rho \) compared with humans could be consistent solely with a larger effective population size in blackbirds versus other species. The data for blackbirds do indeed suggest that \( N_e \) is higher than in humans: Ball et al. (1988) estimated \( N_e \) from mtDNA at about 50,000 and the values for \( \theta_w \) for neutral loci in this study (0.002–0.012) imply an \( N_e \) of ~400,000 if we assume an autosomal non-coding mutation rate of \( 3.6 \times 10^{-9} \) as in gamebirds (Axelsson et al., 2004). However, even these values for \( N_e \) in blackbirds cannot entirely explain the difference in \( \rho \) between blackbirds and humans. We suspect that both \( N_e \) and \( c \) are contributing to the higher \( \rho \) in blackbirds.

(iii) Selection and linkage disequilibrium

Why is the estimate of \( \rho \) at Agph-DAB1 so low relative to the neutral loci in this study? The programs used to estimate \( \rho \) all assume selective neutrality. Although the large amount of polymorphism available in Agph-DAB1 facilitates estimation of \( \rho \), the mode of selection on Mhc loci could yield spurious results with respect to estimating \( \rho \). Epistatic interactions undoubtedly play a role in modulating Mhc allele frequencies, both at the level of linked genes as well as among sets of amino acids linked within the PBR (Hedrick & Thomson, 1983; Hedrick et al., 1991; Hunt et al., 1993; Gaudieri et al., 1999, 2000; Jacob et al., 2000; Jeffreys et al., 2001; Walsh et al., 2003).

In particular, if epistatic interactions among PBR codons are strong, recombination rates could still be high but would not be evident in the standing patterns of diversity because haplotype blocks would be preserved (Ohta, 1999). Assuming the low estimate of \( \rho \) at Agph-DAB1 is real, however, this argument is contradicted by the fact that the stronger block pattern of variation found in mammals, presumably driven by epistasis, is observed despite higher estimated rates of recombination. It is conceivable that stronger selection for polymorphism could be occurring in mammals; this would actually increase the amount of recombination, if recombinant alleles are more likely than \( de\ novo \) mutations to result in new allelic specificities. This argument has the virtue of being consistent with the observation of more recombination at MHC loci in mammals, and suggests the need to obtain precise estimates of selection intensities at avian loci in the future.

Slatkin (2000) has recently modelled the joint effects of selection, mutation and recombination, with the goal of describing the evolutionary forces acting on two-locus haplotypes. He points out that the standard descriptions of linkage disequilibrium may not be adequate to capture fully the dynamics acting on such systems. Because most methods of estimating recombination rely in some form on the frequencies of gametic phase haplotypes, it is likely that new approaches will be required to estimate \( \rho \) when selection cannot be ignored. It may be possible in the near future to estimate jointly selection intensities, recombination rates and population sizes from sequence data (J. Felsenstein, pers. comm.), but it is unclear how well simple selection models will approximate the forces acting on real data. In addition, demographic forces, such as population structure, can strongly influence the pattern of linkage disequilibrium, and hence, estimation of recombination rates (Navarro & Barton, 2002; Wakeley & Lessard, 2003). Population structure and the different genetic backgrounds it produces generally increase the effects of hitchhiking, increasing neutral variability in the vicinity of selected loci, although the precise consequences depend on a number of factors.

Finally, it is also unclear how much data will be necessary to estimate all these parameters jointly with precision from sequence data. For outbred species such as birds that do not yet have or are not amenable to having genetic maps, such estimation will remain a challenge, because several basic parameters available for model species, such as effective population sizes and local recombination rates, are becoming reasonably well known and do not require independent estimation with each study. Nonetheless, the sequence data we have collected in blackbirds strongly indicate that recombination rates in birds
may be substantially higher than in humans, and that
the hitchhiking effect, at least for this particular
.genomic region in blackbirds, may be negligible.
Recent estimates of recombination rates at neutral
loci in *Drosophila* (Wakeley & Hey, 1997; Andolfatto & Przeworski, 2000; Kliman et al., 2000; Hey & Kliman, 2002; Machado et al., 2002) are very high
and suggest that $\rho/\theta$ frequently exceed 1. Thus black-
bird parameters seem to resemble those of *Drosophila
and Peromyscus* more than those of humans. What
is needed now is more nuclear DNA sequence data
from loci known to be linked, in both neutral and
selective contexts. Genomic approaches that allow
large amounts of sequence data to be collected from
multiple nuclear loci will undoubtedly play an im-
portant role in this endeavour (Stephens et al., 2001a;
Clark et al., 2003). Although issues of recombination
and hitchhiking will remain a challenge for clades
such as birds, the consequences of these forces
for topics for which birds are well suited will not
diminish.

We are very grateful to R. Tinghitella for sequencing
the *BACT2* locus; to N. Rotzel for submitting the SNP and
sequence data to the data bases; and to T. Hsieh and M.
Dillon for assistance with Fig. 2. We thank B. Charlesworth,
P. Awadalla, M. Nachman, B. Payseur, J. Felsenstein,
D. Garrigan, S. Schaeffer and A. DiRienzo for helpful
discussion. M. Kuhner, M. Stephens, S. Schaeffer and
G. McVean generously assisted with computer programs.
We thank W. G. Hill for editorial assistance and
B. Jennings, A. Shedlock, H. Walsh, C. Welch and two
anonymous reviewers provided very helpful comments on
the manuscript. This research was supported by grants from
the US National Science Foundation.

Appendix. Inferred haplotypes from Red-Winged Blackbird for the loci used in this study

The set of individuals sampled for the $\beta$-actin locus only partially overlaps those for the other loci. Localities are
as follows: FL, Florida; NIC, Nicaragua; CO, Colorado; OR, Oregon; WA, Washington state; SE, south-east
USA (Florida/Kentucky); NY, New York. Subscript numbers indicate the level of significance of inferred phase
at that site from a typical run of PHASE when analysed simultaneously, as follows: 5, <0.61; 6, 0.61–0.70; 7,
0.71–0.80; 8, 0.81–0.90; 9, 0.91–0.99. No subscript indicates certainty of phase or homozygosity.

(A)

<table>
<thead>
<tr>
<th>Individual no.</th>
<th>Locality</th>
<th>Haplotype</th>
<th>40 kb locus sites</th>
<th><em>Agph-DAB3</em> exon 2</th>
<th><em>Agph-DAB1</em> promoter</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>FL</td>
<td>G G T A G</td>
<td>76 94 135 155 205 206 225 263 301 321 330 337 397 478 15875 15904 40868 40900 40911</td>
<td>15875</td>
<td>15904</td>
</tr>
<tr>
<td>9</td>
<td>NIC</td>
<td>G G T A G</td>
<td>1004</td>
<td>15</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>CO</td>
<td>G G T A G</td>
<td>2005</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>11</td>
<td>OR</td>
<td>G G T A G</td>
<td>243</td>
<td>9</td>
<td>17</td>
</tr>
</tbody>
</table>

We are very grateful to R. Tinghitella for sequencing the *BACT2* locus; to N. Rotzel for submitting the SNP and sequence data to the data bases; and to T. Hsieh and M. Dillon for assistance with Fig. 2. We thank B. Charlesworth, P. Awadalla, M. Nachman, B. Payseur, J. Felsenstein, D. Garrigan, S. Schaeffer and A. DiRienzo for helpful discussion. M. Kuhner, M. Stephens, S. Schaeffer and G. McVean generously assisted with computer programs. We thank W. G. Hill for editorial assistance and B. Jennings, A. Shedlock, H. Walsh, C. Welch and two anonymous reviewers provided very helpful comments on the manuscript. This research was supported by grants from the US National Science Foundation.
### (B) β-fibrinogen intron 7/rhodopsin intron 2

<table>
<thead>
<tr>
<th>Individual no.</th>
<th>Haplotype</th>
<th>BFIB7 sites</th>
<th>RDP2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>49 51 71 108 113 152 204 210 248 275 276 332 344 353 361 387 399 66 175</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>A C G G C A A T A T C T T A A A T A G G</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>A C G G C A A T G T A G C C A A T G T A G G</td>
<td></td>
</tr>
<tr>
<td>1004</td>
<td>7</td>
<td>A C G G C T A T A G C T A G A T A A G G</td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>2</td>
<td>A C C A A A T A T C T A A A T G T</td>
<td></td>
</tr>
<tr>
<td>2020</td>
<td>9</td>
<td>A A C A T A T C T A A A C T T T A A A A T A A T</td>
<td></td>
</tr>
<tr>
<td>2021</td>
<td>5</td>
<td>A C G T C A A T G T A G C C A A G T A A G A</td>
<td></td>
</tr>
<tr>
<td>2022</td>
<td>11</td>
<td>A C A G C A A T G T A G C C A A A T T G A A A A T A A G G</td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>9</td>
<td>A C G T C A A A T G T A G C C A A A T A A G G</td>
<td></td>
</tr>
</tbody>
</table>

### (C) β-actin intron 2

<table>
<thead>
<tr>
<th>Individual no.</th>
<th>Locality</th>
<th>Haplotype</th>
<th>BACT2 sites</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>37 55 104 119 176 215 272 304</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>WA</td>
<td>3</td>
<td>C G G G T A G A</td>
</tr>
<tr>
<td>2</td>
<td>WA</td>
<td>1</td>
<td>C G G G G C G G</td>
</tr>
<tr>
<td>3</td>
<td>WA</td>
<td>3</td>
<td>C G G G T G G G</td>
</tr>
<tr>
<td>4</td>
<td>WA</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>5</td>
<td>WA</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>6</td>
<td>WA</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>7</td>
<td>SE</td>
<td>4</td>
<td>C G G G T A G A</td>
</tr>
<tr>
<td>8</td>
<td>SE</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>9</td>
<td>SE</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>10</td>
<td>NY</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>11</td>
<td>NY</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>12</td>
<td>NY</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>13</td>
<td>SE</td>
<td>3</td>
<td>C G G G T G G G</td>
</tr>
<tr>
<td>14</td>
<td>SE</td>
<td>2</td>
<td>C G G G C G G</td>
</tr>
<tr>
<td>15</td>
<td>SE</td>
<td>1</td>
<td>C G G G C G G</td>
</tr>
</tbody>
</table>

---

*S. V. Edwards and M. Dillon*
References


