MOLECULAR ECOLOGY
RESOURCES

Molecular Ecology Resources (2014) **14**, 1248–1260 doi: 10.1111/1755-0998.12270

Estimation of linkage disequilibrium and interspecific gene flow in *Ficedula* flycatchers by a newly developed 50k single-nucleotide polymorphism array

TAKESHI KAWAKAMI,* NICLAS BACKSTRÖM,* RETO BURRI,* ARILD HUSBY,†‡ PALL OLASON,*[1] AMBER M. RICE,†[2] MURIELLE ÅLUND,† ANNA QVARNSTRÖM† and HANS ELLEGREN*

**Department of Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden, †Department of Animal Ecology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden, ‡Centre for Biodiversity Dynamics, Norwegian University of Science and Technology, N-7491 Trondheim, Norway*

Abstract

With the access to draft genome sequence assemblies and whole-genome resequencing data from population samples, molecular ecology studies will be able to take truly genome-wide approaches. This now applies to an avian model system in ecological and evolutionary research: Old World flycatchers of the genus *Ficedula*, for which we recently obtained a 1.1 Gb collared flycatcher genome assembly and identified 13 million single-nucleotide polymorphism (SNP)s in population resequencing of this species and its sister species, pied flycatcher. Here, we developed a custom 50K Illumina iSelect flycatcher SNP array with markers covering 30 autosomes and the Z chromosome. Using a number of selection criteria for inclusion in the array, both genotyping success rate and polymorphism information content (mean marker heterozygosity = 0.41) were high. We used the array to assess linkage disequilibrium (LD) and hybridization in flycatchers. Linkage disequilibrium declined quickly to the background level at an average distance of 17 kb, but the extent of LD varied markedly within the genome and was more than 10-fold higher in 'genomic islands' of differentiation than in the rest of the genome. Genetic ancestry analysis identified 33 F_1 hybrids but no later-generation hybrids from sympatric populations of collared flycatchers and pied flycatchers, contradicting earlier reports of backcrosses identified from much fewer number of markers. With an estimated divergence time as recently as <1 Ma, this suggests strong selection against F_1 hybrids and unusually rapid evolution of reproductive incompatibility in an avian system.

Keywords: collared flycatcher, GWAS, Hybridization, pied flycatcher, single-nucleotide polymorphism genotyping, speciation

Received 21 January 2014; revision received 14 April 2014; accepted 25 April 2014

Introduction

Genome-wide genetic analyses of populations (population genomics) offer new possibilities for investigating the evolutionary and ecological processes that affect genetic diversity within as well as differentiation between populations. The state-of-the-art currently involves two main routes towards large-scale assessment of the genetic diversity of natural populations. First, the rapid increase in the availability of genome sequences from nonmodel organisms (Ellegren 2014) implies that whole-genome resequencing in population samples is becoming an increasingly realistic means to near-complete genotyping of individual genomes. So far, there are reports of up to tens of resequenced individuals (e.g. Ellegren *et al.* 2012; Jones *et al.* 2012; Varshney *et al.* 2013; Zhao *et al.* 2013); however, sequencing of reduced fractions of the genome has for several years allowed efficient genotyping-by-sequencing of much larger samples (e.g. Gompert *et al.* 2012; Keller *et al.* 2013; Parchman *et al.* 2013). Second, high-density single-nucleotide polymorphism (SNP) arrays represent useful resources for genome-wide genotyping of very large population samples (e.g. Karlsson *et al.* 2011; Wang *et al.* 2011; Van Bers *et al.* 2012; Alhaddad *et al.* 2013; Bourret *et al.* 2013;

Correspondence: Hans Ellegren, Fax: +46-184716310;
E-mail: Hans.Ellegren@ebc.uu.se

[1] Present address: Wallenberg Advanced Bioinformatics Infrastructure (WABI) Science for Life Lab, Uppsala University, Husargatan 3, SE-751 23 Uppsala, Sweden

[2] Present address: Department of Biological sciences, Lehigh University 111 Research Drive, Bethlehem PA 18014, USA

Hagen *et al.* 2013). Analysis of SNP array genotype data is typically straightforward (in comparison with the steps of read mapping, variant calling and genotyping in genomic resequencing), although resolution is by necessity lower than those of whole-genome resequencing. Development of SNP arrays relies initially on SNP discovery, which can be made from a variety of sources, including transcriptome and RAD sequencing (Davey *et al.* 2011). Ideally, with access to an already assembled genome sequence coupled with whole-genome resequencing data, SNPs can be identified and purposely selected for inclusion in an array to cover the entire genome. This study will report on such an endeavour and its downstream application in analyses of linkage disequilibrium (LD) and hybridization in an avian ecological model system.

Genome-wide genotyping in extensive population samples is particularly useful when investigating linkage between specific regions of the genome and trait variation and when characterizing genetic ancestry of individuals at specific genomic regions in the face of gene flow by hybridization. These methods have a high potential of increasing our understanding of the processes of adaptation and the genetics of reproductive isolation between closely related species. In addition, genome-wide population genotyping allows the characterization of the size and distribution of linkage disequilibrium (LD) blocks, which provides important information about how selection moulds the genomic landscape of divergence and diversity (Hohenlohe *et al.* 2012).

Old World flycatchers of the genus *Ficedula* provide a suitable model system for exploring questions related to adaptation and speciation. The collared flycatcher (*Ficedula albicollis*) and the pied flycatcher (*F. hypoleuca*) are small migratory passerine birds that have been subject to long-term research on behaviour and ecology Together, their breeding ranges occupy most of Europe with a broad hybrid zone in central and eastern Europe, most likely established by secondary contact after post-glacial expansion (reviewed in Sætre & Sæther 2010). A much younger hybrid zone, probably only a few hundred years old, is present on the Baltic Sea islands of Öland and Gotland (Qvarnström & Bailey 2009; Lundberg & Alatalo 2010). Other hybrid zones are also likely to have existed during former interglacial periods. The two species hybridize in sympatry at low frequencies (<5%), where the precise rate varies between areas and is related to population densities of the respective parental species (Alatalo *et al.* 1982, 1990; Sætre *et al.* 1997; Veen *et al.* 2001; Borge *et al.* 2005; Svedin *et al.* 2008; Wiley *et al.* 2009). Demographic analyses including approximate Bayesian computation modelling using genome-wide sequence data suggest that the two lineages diverged <1 million years ago (Ma) and that there have been low–moderate levels of almost unidirectional gene introgression from the pied flycatcher to the collared flycatcher (Backström *et al.* 2013; Nadachowska-Brzyska *et al.* 2013).

An outstanding question to speciation research in general and to *Ficedula* flycatchers in particular concerns the relative importance of selection, gene flow by hybridization and demography in shaping genomic divergence. It is theoretically possible that ecologically divergent selection drives species divergence and establishment of reproductive isolation in the face of gene flow (Gavrilets 2004). This may result in a heterogeneous landscape of genomic differentiation with elevated divergence at loci underlying selection. Introgressive hybridization after secondary contact can also result in a heterogeneous landscape by eroding previously accumulated genetic differences with a varying degree depending on variation in recombination rate across a genome (Turner *et al.* 2005; Nosil *et al.* 2009). Recently, we sequenced and assembled the 1.1 Gb genome of collared flycatcher and identified 13 million SNPs by resequencing 10 individuals each of collared flycatcher and pied flycatcher. Using these data, we identified ~50 distinct 'genomic islands' with elevated divergence between the two species by Ellegren *et al.* 2012. To understand the causal processes and mechanisms behind such islands, the possibility to perform genome-wide genotyping approaches on large number of individuals would be important.

Here, we describe the development of a 50K Illumina iSelect SNP array for fine-scale genomic analysis of *Ficedula* flycatchers. With the aid of this array, we investigate the levels of LD in a collared flycatcher population, a parameter which is not only informative for population genetic processes but also will affect the prospects for successful GWAS efforts. Moreover, we address the case for ongoing gene flow between collared flycatcher and pied flycatcher by array genotyping of population samples as well as of putative hybrids.

Materials and methods

Specimens

Blood samples were collected from 422 collared flycatchers and 59 pied flycatchers breeding in sympatry on the Baltic Sea island Öland from 2002 to 2011 (Table 1). According to the detailed pedigree information for these populations (Svedin *et al.* 2008; A. Qvarnström, unpublished data), the samples represented putatively unrelated individuals without direct descendants. We also used blood samples of 48 putative hybrids, which were either offspring in nests of breeding mixed species pairs or adults identified by their intermediate plumage characters (typically broken collar on their neck) or mixed

Table 1 Flycatcher samples used in this study

Species	Total	Male	Female
Collared flycatchers	422	221	201
Pied flycatchers	59	35	24
Hybrids identified based on phenotypic characters	31	24	7
Hybrids in nests with mixed breeding pairs (CF × PF)*	2	1	1
Hybrids in nests with mixed breeding pairs (PF × CF)*	9	7	2
Hybrids in nests with mixed breeding pairs (HY × CF)*	6	3	3

*Putative hybrid individuals found in a mixed breeding nest (sire × dam, CF, collard flycatcher; PF, pied flycatcher; HY, hybrid).

song (incorporation of components from both parents' songs) (Table 1). DNA was extracted by a standard proteinase K digestion/phenol-chloroform purification protocol (Sambrook *et al.* 1989).

Single-nucleotide polymorphism array design

We recently reported a reference collared flycatcher genome (~1.1 Gb in size) by sequencing one male bird at 85X coverage (Ellegren *et al.* 2012). In addition, we found more than 13 million segregating SNPs by mapping short sequence reads from 9 collared flycatchers and 10 pied flycatchers (mean genomic sequence coverage per individual and site was 5.69X) to the reference collared flycatcher genome. These SNPs include sites polymorphic in one or both of the species and sites fixed for different alleles in the two species. The latter represent a clear minority of sites given that lineage sorting is far from complete in these two recently (<1 Ma) diverged species (Nadachowska-Brzyska *et al.* 2013).

As the SNP array was primarily designed for subsequent use in collared flycatcher populations, we focused on selecting markers found to be polymorphic in this species based on the abovementioned resequencing data. As a starting point for selecting SNPs for inclusion in the array, we maximized the number of markers by only considering variants that required a single probe assay in the Illumina Infinium II genotyping system (Illumina Inc., San Diego, CA) (i.e. A/C, A/G, C/T and G/T polymorphisms; Gunderson 2009). We then ranked all SNPs by the following six criteria to maximize the usefulness of the array. First, we chose markers that were heterozygous in the individual used for genome assembly (where the high coverage should imply a negligible rate of falsely called SNPs) and/or polymorphic in an independent set of individuals used in transcriptome sequencing by Ellegren *et al.* (2012). Second, markers present in coding regions of genes, based on MAKER (Cantarel *et al.* 2008) gene predictions, were preferred because these markers may be more likely to show linkage to causative variants for traits of interests than markers in intergenic regions. Third, for scaffolds >25 kb in size in a preliminary genome assembly version present at the stage of array design, we ensured to have at least two markers per scaffold to facilitate ordering and orienting scaffolds along chromosomes if the array would be used for pedigree genotyping and linkage mapping in the future. Fourth, markers with as high minor allele frequency (MAF) as possible were preferred. Fifth, markers with high Genome Analysis Toolkit (GATK) (DePristo *et al.* 2011) variant quality scores were preferred. Sixth, we sought to obtain an even distribution of markers across the genome in terms of putative genetic distances. In the absence of a high-density flycatcher linkage map, we used a zebra finch linkage map (Stapley *et al.* 2008) as a reference to estimate genetic distance. To this end, LASTZ (Harris 2007) was used to align the collared flycatcher and zebra finch genomes, and the alignments were split into 0.5 cM bins according to the zebra finch linkage map. Of a total of 3852 bins, 3550 bins included markers from our SNP list.

We started the process of marker selection by choosing the SNPs that fulfilled all criteria and ranked highest in terms of polymorphism and quality. When there were no further markers fulfilling all criteria, the criteria were gradually relaxed. Furthermore, we iteratively updated the marker list based on the evaluated Illumina design score, which measures the quality of the probe sequence of the variant to be genotyped. Variants with failing assay scores (threshold score = 0.6) were replaced until all submitted variants passed the design score filtering check. Importantly, we ensured that probe sequences would not cover polymorphic sites, as this would lead to

Table 2 Summary of marker polymorphism for informative SNPs included in the array

Marker type	Category I[1]	Category II[2]	Total
Polymorphic in both species	14 944	1112	16 056
Polymorphic in collared flycatcher	21 196	1837	23 033
Polymorphic in pied flycatcher	19	454	473
Fixed between species	136	829	965
Invariant	352	14	366
Total	36 647	4246	40 893

[1] Polymorphic markers in the panel individuals of collared flycatchers (n = 10).

[2] Fixed markers with alternate alleles between the panel individuals of collared and pied flycatchers (n = 10 each).

the risk of null alleles. In total, we selected 45 000 markers polymorphic in collared flycatchers, and these markers will be referred to as Marker Category I (Table 2). In addition, we selected 5000 markers that were fixed with alternate alleles in our panels of 10 collared flycatchers, including the bird used for the genome assembly, and 10 pied flycatchers (Marker Category II). Although the size of the panels was not sufficiently large to expect that these 5000 markers would necessarily be fixed for alternate alleles in larger population samples, we anticipated that a significant proportion of them would be informative for hybrid detection. Finally, all markers were submitted to make a custom SNP array using the iSelect BeadChip (Illumina).

Genotyping and characterization of linkage disequilibrium

Genotyping was performed on an Illumina iScan instrument at the SNP & Seq Technology Platform at Uppsala University (http://www.molmed.medsci.uu.se/SNP+SEQ+Technology+Platform/). SNPs that resulted in more than three genotype clusters (i.e. heterozygote and two alternate homozygote genotypes) were removed from subsequent analysis.

Pairwise linkage disequilibrium in the form of r^2 was calculated for collared flycatcher using Haploview version 4.2 (Barrett *et al.* 2005) for pairs of SNPs that were uniquely mapped onto the genome and had MAF >10% and genotyping rate >90%. To estimate physical distances between markers located in separate scaffolds, scaffolds were concatenated into chromosome sequences with an arbitrary gap size of 5 kb. The decay in LD over distance was estimated for each chromosome separately or all chromosomes together by nonlinear regression for pairs of SNPs within and outside 'genomic islands' of differentiation identified by (Ellegren *et al.* 2012). Briefly, these were defined as genomic regions where at least two adjacent 200 kb windows had a density of fixed differences between collared flycatcher and pied flycatcher exceeding 0.001 per bp. The expected value of r^2, $\mathrm{E}(r^2)$, under drift-recombination equilibrium can be expressed as (Hill & Weir 1988):

$$\mathrm{E}\left(r^2\right) = \left[\frac{10+\rho}{(2+\rho)(11+\rho)}\right]\left[1+\frac{(3+\rho)\left(12+12\rho+\rho^2\right)}{n\,(2+\rho)\,(11+\rho)}\right]$$

where ρ is the population recombination parameter ($\rho = 4N_\mathrm{e}r$) and n is the number of sequences sampled. Linear models were used to test whether the variation in chromosome size was correlated with estimated ρ. Chromosome 22 and linkage group LGE22 were excluded from this analysis because of the small number of markers on these chromosomes (45 and 36 SNPs, respectively). Chromosome Z was also excluded due to the special character of sex chromosome recombination. Box–Cox power transformations were used to normalize the distribution of the predictor variable (i.e. chromosome size).

Tag SNPs are SNPs that represent the genetic variation in a region, selected based on the degree of linkage disequilibrium with adjacent SNPs (Johnson *et al.* 2001) that would characterize the full set of markers on the array by running the tagger analysis in Haploview (de Bakker *et al.* 2005; Barrett *et al.* 2005) with default settings for minimum linkage between SNPs at threshold LD $r^2 = 0.5$ or 0.8. Block structure within chromosomes was assessed by calculating block sizes for all instances where at least two adjacent SNPs showed lack of evidence for recombination in the sample (i.e. the four-gamete test).

Analysis of hybrids

To estimate genetic ancestry of the 48 putative hybrids, we used the maximum-likelihood-based clustering approach implemented in ADMIXTURE 1.04 (Alexander *et al.* 2009). To reduce redundant information coming from closely linked markers, we pruned the data set using Plink (Purcell *et al.* 2007; default settings) by excluding SNPs that had LD $r^2 > 0.1$ with other SNPs within 50-SNP overlapping sliding windows (advanced by 10 SNPs). Z-linked markers were also removed from the analysis. Default input parameters were used to estimate ancestry fractions Q (block relaxation optimization, unsupervised learning option, secant condition parameter $q = 3$). In addition, to distinguish among different classes of hybrids (F_1's, F_2's and backcrosses), we jointly estimated hybrid index (S) and interspecific heterozygosity (H_I) of these individuals using HIest with Markov chain Monte Carlo option with 1000 iterations (Fitzpatrick 2012). For this analysis, we used markers that were (i) mapped on autosomes; (ii) fixed with alternate alleles between our sample cohorts of collared flycatchers and pied flycatchers; and (iii) separated from each other with >100 kb to minimize redundant information resulting from linkage between markers.

To compare allele frequency distributions between collared flycatchers and F_1-hybrids, allele frequencies in the hybrids were calculated for markers that were invariant in pied flycatchers by subtracting alleles contributed by pied flycatcher chromosomes (hereafter, referred to as 'pied-fixed alleles'). Markers with too many missing genotypes (call rate <90% of individuals in each population) were excluded. Allele frequencies of Z-linked markers were calculated for males only. Over-/under-representation of pied-fixed alleles at each locus in the hybrid population was tested by Fisher's exact tests after applying a 5% false discovery rate (FDR) for multiple testing (Benjamini & Hochberg 1995).

Results

Single-nucleotide polymorphism array performance

Of 50 000 selected markers, the final array contained 45 138 markers, which corresponds to a 90% success rate in array construction. Forty-one thousand hundred and sixty-seven of the successfully printed markers were uniquely assigned to one of the chromosomes of the collared flycatcher assembly (version FicAlb_1.4; Ellegren *et al.* 2012), while the physical position of the remaining 3971 markers was unknown (Fig. 1). As we sought to select markers with an even coverage of the genetic map, and given that the recombination landscape of avian genomes appears highly heterogeneous, marker density varied considerably within and among chromosomes (Fig. 1). The density was higher for small chromosomes than for large chromosomes and higher towards the ends of chromosomes than in central regions, a consequence of the higher expected recombination rates in these regions (Stapley *et al.* 2010, Backström *et al.* 2010).

We used the array for genotyping of sympatric population samples consisting of 422 collared flycatchers and 59 pied flycatchers. The initial quality filtering removed 2083 SNPs because they formed more than three genotype clusters. In addition, 2162 SNPs produced low scorable genotypes in collared flycatcher samples (<5% of total samples), and these were also removed from the subsequent analyses. For the remaining 40 893 markers (Table 2), the genotype call rate was >98%, and reproducibility among duplicate analysis of 50 samples was 100%.

Figure 2 summarizes the overall degree of polymorphism for markers from Category I (selected for being polymorphic in collared flycatchers) and II (selected for being fixed for alternate alleles in resequencing of 10 birds of each species) in collared flycatchers and pied flycatchers (Table 2). The distribution of MAFs for Category I markers in collared flycatcher clearly shows the benefit of selecting markers based on an initial assessment of polymorphism using whole-genome resequencing; the distribution was highly skewed towards high MAFs (Fig. 2a), in contrast to the expected skew towards rare variants for a random set of segregating sites. About 40% of markers polymorphic in collared flycatchers were also polymorphic in pied flycatchers (Fig. 2b). Overall

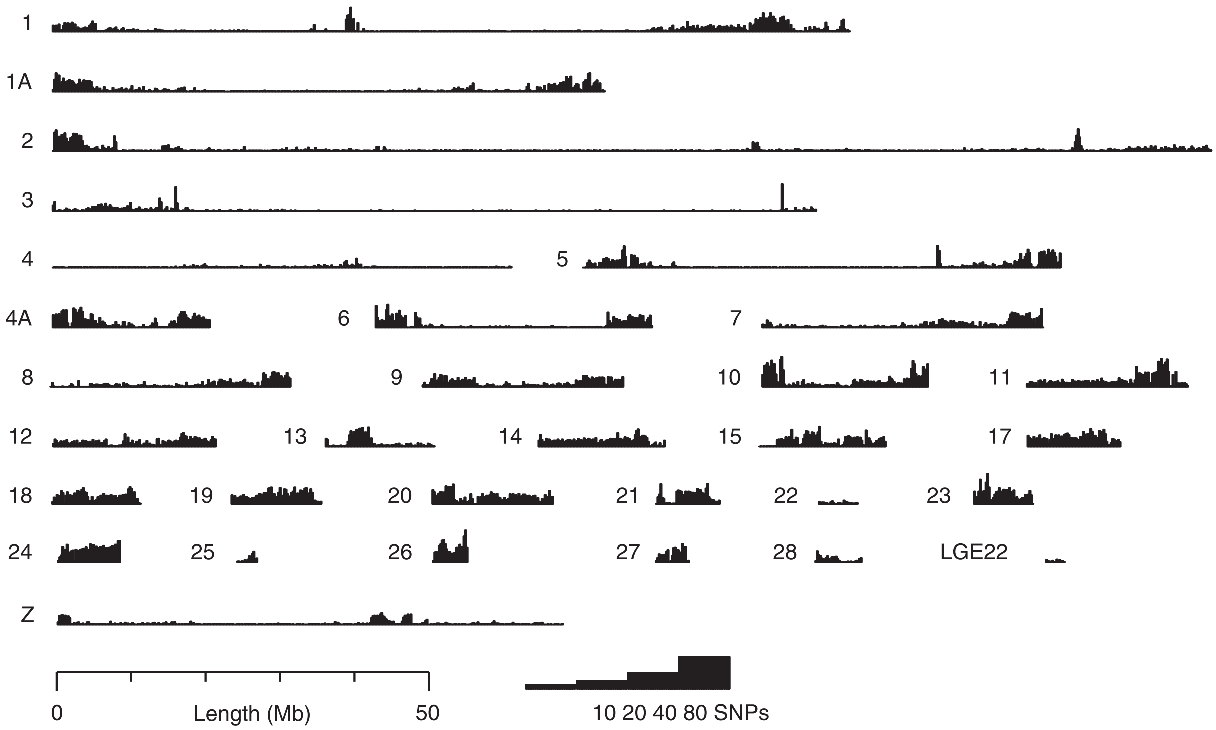

Fig. 1 Distribution of 45 138 markers presents on a custom collared flycatcher single-nucleotide polymorphism array. The number of markers in 200 kb bins is shown along chromosomes in the collared flycatcher genome (Ellegren *et al.* 2012). Note that 3971 markers with unknown genomic locations are not shown.

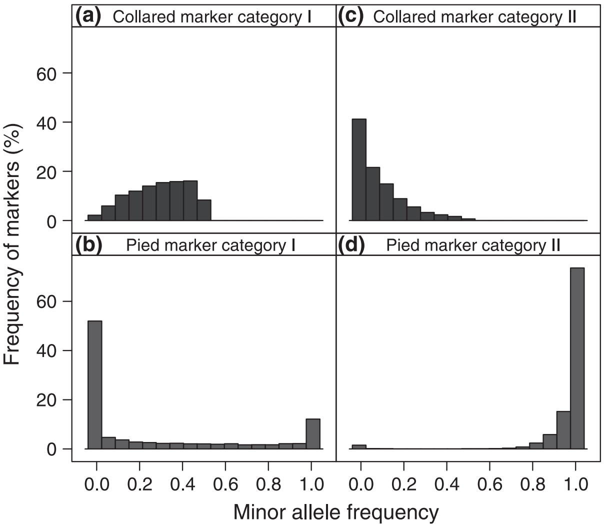

Fig. 2 Distribution of minor allele frequency (MAF) for Category I (a and b) and Category II (c and d) markers in collared flycatcher and pied flycatcher, respectively. Category I includes 36 647 markers that are selected for being polymorphic in collared flycatchers, while Category II includes 4246 markers that were selected for being fixed for alternate alleles in resequencing of 10 collared flycatchers and 10 pied flycatchers.

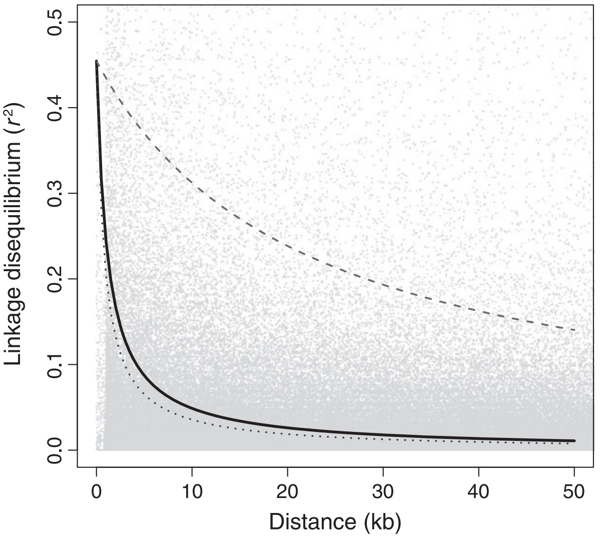

Fig. 3 Plot of linkage disequilibrium (LD, r^2) against distance between SNPs in collared flycatcher. Grey dots indicate observed pairwise LD. Solid black, dashed red and dotted blue curves show the expected decay of LD in the genome-wide data, as well as within and outside the divergence regions, respectively, estimated by nonlinear regression of r^2.

heterozygosity for the Category I markers was 0.415 and 0.172 in collared flycatcher and pied flycatcher, respectively. Mean folded MAFs for these markers were 0.289 and 0.086 in collared flycatcher and pied flycatcher, respectively. For markers from Category II, assumed to be enriched for diagnostic sites, 19.5% of markers were fixed with alternate alleles between populations of the two species (Fig. 2c,d). This demonstrates that low- to medium-coverage resequencing of a small number of individuals of each species had a relatively low success rate in detecting sites where the two species are fixed for different alleles.

Linkage disequilibrium

Overall LD decayed rapidly and reached the background level (average r^2 = 0.031) at an average distance of about 17 kb between markers (Fig. 3). Consistent with the genome-wide pattern of rapid decay in LD, the majority of LD blocks were short; median block length was 3.0 kb, and mean block length was 20.5 kb (Fig. 4). The longest LD blocks were found on chromosome 4 (1675 kb), chromosome 2 (1626 kb) and chromosome 5 (1556 kb); however, these blocks contained only two SNPs and might thus be inconclusive. The LD blocks with the largest number of linked SNPs were located on chromosome 14 (30 SNPs within a 396 kb block) and chromosome 27 (20 SNPs within a 330 kb block) (Fig. 4). As expected, given the low degree of LD, as many as 32 289 (95.4%) tag SNPs were needed to represent the 33 820 SNPs for a threshold LD of r^2 = 0.5 and 33 101 tag SNPs (97.8%) were needed for r^2 = 0.8. The ratio of tag SNPs to the full set of SNPs was high on all chromosomes (range 92.5–100%). Tag SNP selection statistics are presented in detail in Table S1 (Supporting information).

Our previous work revealed that collared flycatcher-pied flycatcher differentiation is highly heterogeneous across the genome (Ellegren *et al.* 2012). In particular, differentiation in some 50 distinct 'genomic islands' was up to 10-fold higher compared with background levels in the genome. LD extended much further within the islands than outside (average r^2 = 0.137 vs. 0.025 within and outside genomic islands; Wilcoxon test, *P*-value <2.2e-16). The estimated population recombination parameter *C* was two orders of magnitude smaller within the 'divergence regions' than outside (0.093×10^{-3} vs. 2.534×10^{-3}). Consequently, the expected LD dropped to the background level at much larger distance for markers within genomic islands than outside these regions (11.7 and 318.8 kb, respectively).

Hybridization between collared flycatcher and pied flycatcher

The genetic admixture analysis (Alexander *et al.* 2009) using a pruned subset of autosomal markers with LD ≤0.1 with each other (21 186 SNPs) supported the clear genetic subdivision of collard flycatchers and pied

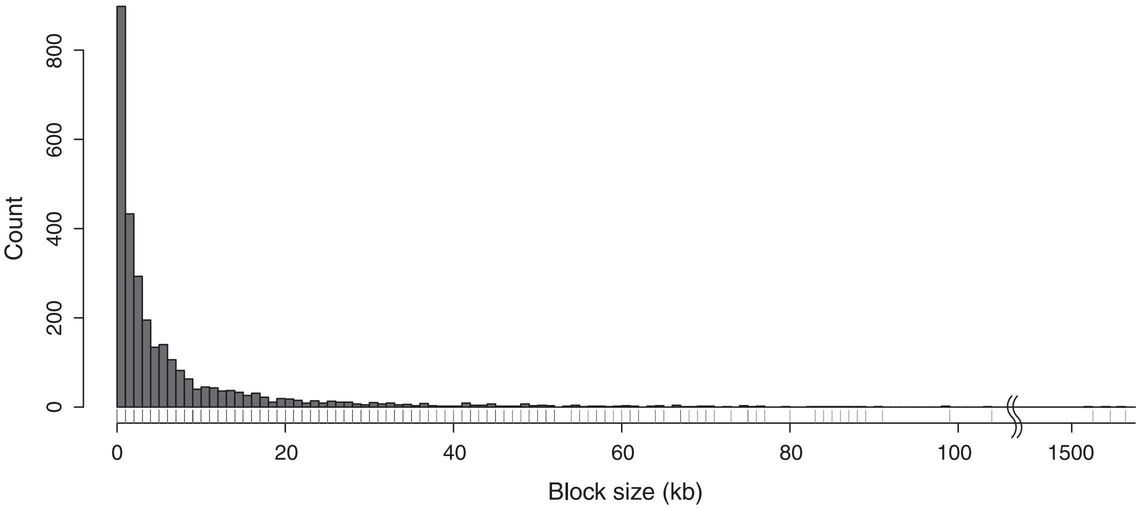

Fig. 4 Distribution of linkage disequilibrium block sizes across all chromosomes. Red vertical bars below the x-axis (rug) illustrate position of bars in the histogram.

flycatchers (Fig. 5). All collared flycatchers except two individuals (i.e. 420 in total) had nearly 100% collared flycatcher genetic ancestry (blue bars in Fig. 5). Similarly, 56 of 59 pied flycatchers had nearly 100% pied flycatcher genetic ancestry (green bars in Fig. 5). However, two collared flycatchers and three pied flycatchers, recorded as 'pure' species upon sampling, had F_1-like genotypes with ~50% each of collared flycatcher and pied flycatcher genetic ancestry.

In addition, we genotyped a total of 48 putative hybrids (Table 1). Among 31 birds with intermediate phenotypic characters, 24 individuals (20 males, 4 females) had F_1-like genotypes with an equal contribution of the two species in their ancestry. The genetic ancestry of the remaining seven individuals turned out to be either 'pure collared' (four males, two females) or 'pure pied' (one female). For putative hybrids sampled as nestlings from breeding mixed pairs of the two species, some interesting observations were made. There were two mixed breeding pairs comprising collared flycatcher male and pied flycatcher female, and, as expected, their offspring (only one per nest) had F_1-like genetic ancestry (one male and one female). However, for nine offspring sampled in the nests of five mixed pairs comprising pied flycatcher male and collard flycatcher female, only two of them had F_1-like genetic ancestry. The remaining seven offspring had 'pure collared'-like genotypes (five males, two females),

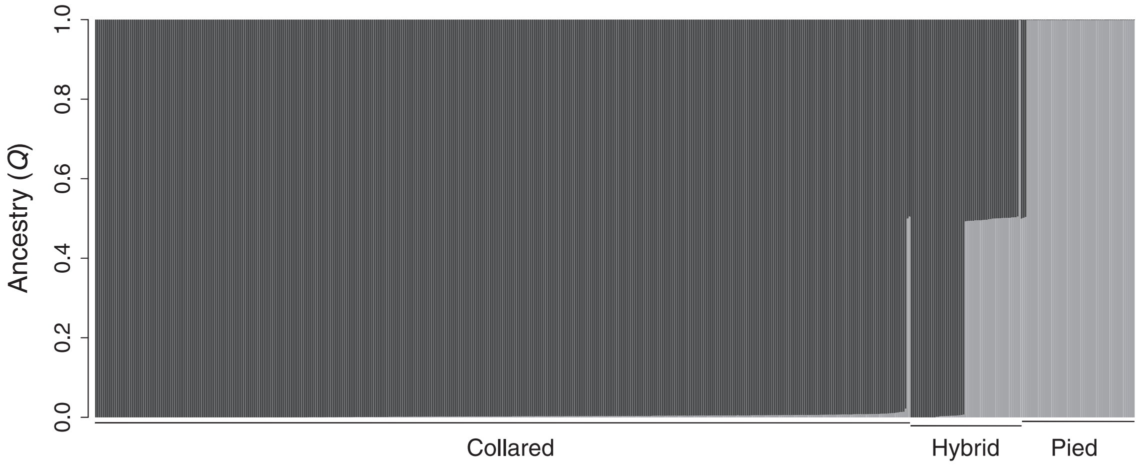

Fig. 5 Genetic ancestry analysis of sympatric populations of collared flycatcher (422 individuals), pied flycatcher (59 individuals) and their putative hybrids (48 individuals) using ADMIXTURE. Genetic ancestry clusters are indicated with blue for collared flycatchers and green for pied flycatchers.

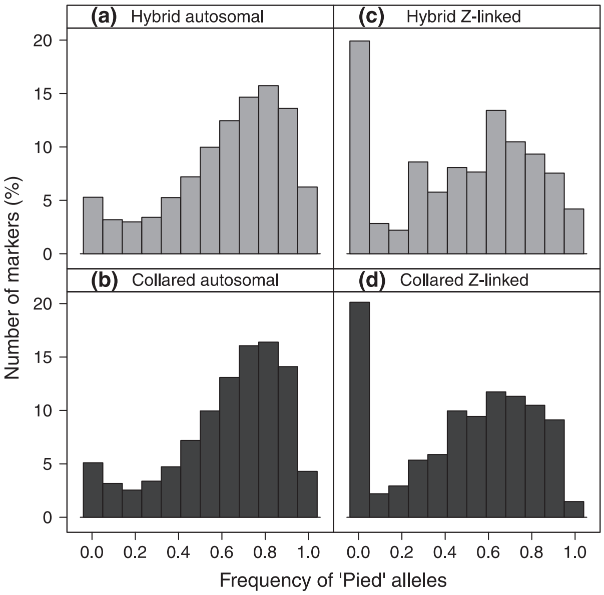

Fig. 6 Allele frequency distribution of collared (blue bars) and 33 F_1-like hybrids (grey bars) at 23 846 markers that were invariant in pied flycatchers. (a and b, 22 883 autosomal markers; c and d, 963 Z chromosome linked markers). Note that allele frequencies in the hybrid population were calculated by subtracting alleles contributed by pied flycatcher parents.

suggesting that their biological fathers were not the same individuals as the social father (i.e. extra-pair paternity; however, in none of these cases were the social father genotyped so we could not confirm this). Finally, six offspring were sampled in a nest at which an F_1 hybrid male bred with a collared flycatcher female, suggesting that they would represent backcrosses. However, all six individuals had ~100% collared flycatcher genotypic ancestry, again consistent with extra-pair paternity. We thus found 28 F_1-like hybrids of 48 putative hybrid samples but did not find any backcrosses or more advanced stages of introgression.

In total, our genetic admixture analysis identified 33 F_1-like hybrids, including those identified in screening the supposedly pure population samples. Joint estimates of hybrid index (S) and interspecific heterozygosity (H_I) of these hybrids using 466 fully diagnostic markers (i.e. markers fixed with alternate alleles between the two species) revealed that all of these individuals had indeed F_1 genotypes with hybrid index (S) ≈ 0.5 and interspecific heterozygosity (H_I) ≈ 1 (Fig. S1, Supporting information).

Genome scans of advanced-generation hybrid individuals provide a means for the identification of regions resistant to introgression, indicative of harbouring incompatibility loci. As we only detected F_1 hybrids in array genotyping, this approach is not possible with this sample; the access to individuals of at least the first backcross generation would be needed for this. However, we can still evaluate another possible form of 'biased' introgression: nonrandom introgression of segregating alleles of polymorphic sites in cases when one of the species is fixed for one of these alleles. We focused on 23 846 markers that were invariant in pied flycatchers and polymorphic in collared flycatcher. In this case, we can trace which allele the collared flycatcher parent has transmitted to hybrid offspring and estimate the frequency of either of these alleles among hybrid offspring and compare this with allele frequencies in the collared flycatcher population.

Overall, allele frequency distribution of collared-derived alleles among the 33 F_1-like hybrids was similar to that in the collared flycatcher population: mean frequencies of 'pied-fixed alleles' were 0.6187 and 0.6191 for hybrids and collared flycatchers, respectively (Wilcoxon paired test, $P = 0.595$) (Fig. 6). Nonetheless, four markers showed highly significant under-representation of 'pied-fixed alleles' in the hybrid population (Table 3; these were the only markers significant at $P < 0.05$ after 5% FDR). As an example, at marker *S00053:282810*, pied flycatchers were fixed for an A allele and counts in collared flycatchers were 493 of the A allele and 345 of the alternative G allele. However, among the 33 hybrids, 32 were heterozygous AG and only one was homozygous AA, indicating biased transmission from the collared flycatcher parent in favour of the G allele. These four markers were located at chromosome 1A, 8, 14 and 27. The marker on chromosome 8 was located in the 5′-untranslated region (UTR) of the eukaryotic translation initiation factor 2B subunit 3 γ(*EIF2B3*) gene, while the other three markers were found in intergenic regions. SNPs in the neighbourhood to these markers showed no sign of biased transmission, however, given the low degree of LD this may not be unexpected.

Discussion

Flycatcher single-nucleotide polymorphism array

Recent advancement in high-throughput sequencing technologies has enabled rapid and reliable discovery of genome-wide SNP markers for ecologically important organisms. We developed a custom collard flycatcher 50K SNP array using the comprehensive genomic resources recently developed for *Ficedula* flycatchers in the form of a draft genome assembly, genome-wide SNP discovery by whole-genome resequencing of population samples and transcriptome sequencing (Ellegren *et al.* 2012). The SNP array has markers that cover all 30 characterized chromosomes in the genome assembly plus all large unassigned scaffolds, a proportion of which are

Table 3 List of loci with a significant allele frequency difference between the collared flycatcher population and hybrids. The test included loci at which pied flycatchers were fixed for one of the two alleles segregating in the collared flycatcher population

	Genomic location		Alleles		Frequency of pied-fixed allele		
Marker name	Chromosome	Position (Mb)	Pied-fixed	Alternative	Collared flycatcher	Hybrids	*P*- value*
S00155:375545	1A	61.0	A	G	0.534	0.061	1.15E-04
S00053:282810	8	12.3	A	G	0.588	0.030	4.57E-07
S00038:4584510	14	13.8	G	A	0.619	0.121	6.43E-05
S00199:1006768	27	2.3	G	A	0.510	0.030	5.69E-05

*Fisher's exact test after 5% false discovery rate correction.

likely to originate from still uncharacterized microchromosomes.

The SNP selection criteria that we employed for array development proved highly successful. Only 2162 markers failed to produce scorable genotypes in both collared and pied flycatchers (95% success rate), suggesting that (i) flanking sequences for probe design were successfully extracted from the reference collared flycatcher genome (Ellegren *et al.* 2012) and (ii) these flanking sequences were well conserved between and within species. High reproducibility using replicated samples (50 of 50) of collared flycatchers confirms a reliable genotyping with very low error rate. When it comes to informativeness, two important observations were made. First, by selecting markers based on the degree of polymorphism initially seen in whole-genome resequencing, it is possible to obtain an array with markers of high polymorphism information content. Second, the resequencing of 9–10 individuals of each species was in most cases not sufficient for identifying sites truly fixed for alternate alleles. Of 4246 Category II markers initially suggested to be species-diagnostic, only 19.5% were subsequently found to be fixed for alternate alleles in the much larger population samples (Table 2). However, largely nonoverlapping allele frequency spectra (Fig. 2) still support the utility of Category II markers for charactering patterns of genetic admixture between collared and pied flycatchers (Fig. 5 and Fig. S1, Supporting information).

Linkage disequilibrium decay

We have previously reported that genome-wide divergence between collared flycatcher and pied flycatcher is highly heterogeneous and is represented by ~50 'genomic islands' that are usually associated with higher LD than background genomic regions (Ellegren *et al.* 2012). Our new genotype data using the SNP array confirmed this observation by showing larger mean LD between pairs of markers, and slower LD decay, within the islands than outside these regions (Fig. 3). Variation in the extent of LD is caused by a number of inter-related mechanisms, such as differences in recombination rate, mutation rate, genetic diversity, selection, effective population sizes and genetic drift (Hill & Robertson 1968; Barton 2000; Pritchard & Przeworski 2001; Wang *et al.* 2002; Stumpf & McVean 2003; Rundle & Nosil 2005; Slatkin 2008). As many of the islands appear to be located close to predicted centromeric and telomeric regions (Ellegren *et al.* 2012), extended LD can also be associated with underlying molecular and genetic features at centromeric/telomeric regions. Several species show reduced recombination rate near centromeres (chicken, Groenen *et al.* 2009; tomato, Sherman & Stack 1995), while other species showed that variation in recombination rate was not strongly correlated with centromeres (domestic pig, Tortereau *et al.* 2012). As the location of centromeres was predicted based on the homologous chromosomes of zebra finch, the accurate karyotype of collared flycatcher will be essential for further investigating the relationship between the extent of LD and centromeric/telomeric regions.

We have previously reported rather extended LD in the collared flycatcher using 34 SNPs from 23 different genes on the Z chromosome; LD in the form of D' dropped below 0.5 at ~400 kb (Backström *et al.* 2006). When the same LD metric was used here, D' dropped below 0.5 at ~240 kb using 770 SNPs on the Z chromosome. Shorter LD in the current study may result from much larger sample size (82 females vs. 221 males) as well as higher marker density. In addition, the markers of our previous study span roughly 50% of the entire Z chromosome, while the markers of the current study cover nearly 100% of the Z chromosome. As the ends of avian chromosomes tend to have higher recombination rates than centre of chromosomes (Backström *et al.* 2010), this may have resulted in the detection of a more rapid decay of LD in the present study. Such bias may very well be a general feature of population genetic studies using limited amounts of markers.

The SNP array presented here offers a valuable resource for future studies of *Ficedula* flycatchers, such as linkage mapping, association mapping, LD mapping and scans for selective sweeps. At the same time, our results also highlight the current difficulties and challenges in developing genomic toolkits for natural populations. First and foremost, because of the rapid decay of LD in most parts of the flycatcher genome, completely covering the entire genome with independent SNP sets would require a much larger number of markers. For instance, given the distance over which LD decays to the background level (mean of 17 kb on each side), >32 000 evenly distributed markers with distance between markers ≤34 kb would be required to cover the entire flycatcher genome with the size of 1.1 Gb. As the flycatcher array contained ~13 000 polymorphic markers after pruning markers within 34 kb from neighbouring markers, ~60% of the flycatcher genome is not covered by markers represented on the array. Second, consistent with the above calculation, the tag SNP and block structure analysis revealed that the number of markers on the array is too low to completely cover variation in the whole genome. With a moderate LD threshold of $r^2 = 0.5$, it is still required to use 95.4% of the markers (32 289 tag SNPs) to efficiently represent all markers on the array. Finally, the genome-wide pattern of rapid LD decay is further illustrated by the existence of a large number of short LD blocks with <1 kb and with a median block size of 3.0 kb. However, it should be noted that our set of markers was biased towards high-recombination regions, resulting in the recovery of a large number of small LD blocks.

Hybridization

The collared flycatcher and the pied flycatcher are almost completely reproductively isolated from each other, yet occasionally form heterospecific breeding pairs and hybridize, which creates individuals of mixed ancestry (reviewed by Sætre & Sæther 2010 and references therein). The fitness of hybrids is severely reduced, with apparent sterility of females (Alatalo *et al.* 1990; Gelter & Tegelström 1992) and with reduced fertility of males (Ålund *et al.* 2013) and sexual selection against intermediate phenotypes contributing to reduced male fitness (Svedin *et al.* 2008; see further below). According to field observation, about 4% of breeding pairs are mixed and about 3% have a hybrid male breeding in our sympatric study populations (Svedin *et al.* 2008). The genetic admixture analysis identified a total of five F_1 hybrids in our main sample cohort of collared flycatchers and pied flycatchers (Table 1). It is not surprising to find a small portion of hybrids because of the difficulty in confidently identifying hybrids of these species, in particular of females. This is also reflected in the fact that seven of 31 individuals classified as hybrids based on their phenotypic characters actually turned out to have genotypes corresponding to pure species (Table 1; cf. Veen *et al.* 2001). Interestingly, of 17 offspring from mixed breeding pairs, only four had F_1 hybrid genotypes (21%). Extra-pair paternity occurs relatively frequently in collared flycatchers (Sheldon & Ellegren 1999), and even more so in mixed pairs (Veen *et al.* 2001), and it has been suggested that female collared flycatchers can reduce the indirect costs of mixed pairing (unfit offspring) by engaging in conspecific extra-pair copulations, either as an active strategy or favoured via conspecific sperm precedence (Veen *et al.* 2001). At the same time, direct benefits of hybridization could be accrued via ecological factors, such as habitat conditions (Veen *et al.* 2001; Wiley *et al.* 2007). Our results quantitatively support that conspecific fathers often sire offspring from mixed pairings in this system.

Our SNP array contained 965 fully diagnostic markers to distinguish the two flycatcher species, and a subset of these markers with minimal linkage (466 markers, >100 kb apart from each other) was applied to characterize genetic ancestry of 33 putative hybrids. The ancestry analysis revealed that all of these individuals had F_1 hybrid genotypes, and there were no backcrosses or more advanced later-generation hybrids (Fig. S1, Supporting information). For the same populations, Wiley *et al.* (2009) estimated the incidence of F_1 hybrids, first-generation backcrosses (B_1) and second or later-generation backcrosses (B_2) as 0.9–1.8%, 0.4–0.5%, and <0.3%, respectively, using 40 informative SNPs to distinguish these species. As most of the hybrids detected in the present study did not come from a random sample of birds, the incidence of hybrids is not directly comparable between these two studies. Importantly, however, three hybrids genotyped in both studies were classified as B_1 (one) and B_2 (two) hybrids by Wiley *et al.* 2009, whereas all of them had hybrid F_1 genotypes in the present study. This discrepancy is likely explained by the difference in the number of markers used (40 vs. 466), the number of individuals screened for assessing species-specificity, and/or because the 40 SNPs in the previous study were considered informative based on allele frequency distributions in separate allopatric populations. It could thus be that backcrosses are extremely rare in these sympatric flycatcher populations and that the amount of ongoing gene flow is very low, if at all present. More generally, this illustrates the limitations associated with admixture analyses when genome-wide approaches cannot be taken.

Absence of backcrosses and later-generation hybrids implies strong selection against F$_1$ hybrids. Using approximate Bayesian computation (ABC) for reconstructing the demographic history and timing of speciation, we recently estimated their divergence to be <1 Ma and gene flow from pied flycatcher into collared flycatcher at a rate of 0.16–0.36 migrants per generation (Nadachowska-Brzyska *et al.* 2013). Although the timing of gene flow could not be precisely ascertained, a model with recent gene flow after the last glacial maximum (LGM) was suggested. If this is correct, several scenarios for how to view the present results are possible. One is that the rate of gene flow differs between different hybrid zones and areas of secondary contact of these species in such a way that it is lower in our study populations than elsewhere. Flycatcher populations on the Baltic Sea islands Gotland and Öland have most likely come into secondary contact only recently (Qvarnström *et al.* 2010). However, if anything, one might have expected stronger barriers to gene flow in old hybrid zones than in areas of recent contact. A previous study using 25 microsatellite loci and 20 SNPs supports this scenario, with the highest introgression in populations of Gotland and Öland (Borge *et al.* 2005). Alternatively, it could be that our sampling regime during field studies does not provide a random representation of the population, for example, because hybrids are more dispersive. Of course, given the uncertainty in the ABC estimates, it may be that gene flow only occurred up until, or before, the LGM and that strong reproductive incompatibilities evolved very recently.

Previous studies suggest that various types of mechanisms are involved in the reproductive isolation of these species (Qvarnström *et al.* 2010; Sætre & Sæther 2010). First, mating success rate of hybrid males is lower than pure males of either species because of their intermediate plumage characters and mixed song, which is disadvantageous for attracting mating partners (Svedin *et al.* 2008). Second, even after successful mating, genes of hybrid males were less likely to contribute to the subsequent generations due to the low hatching rate of their offspring and high susceptibility to extra-pair paternity, where hatched nestlings were likely to be sired by other males of the pure species (Svedin *et al.* 2008; Ålund *et al.* 2013). Finally, fitness of F$_1$ hybrids is much lower than pure species due to complete sterility in female hybrids and severely reduced reproductive performance of male hybrids by producing a high proportion of malformed sperm (Alatalo *et al.* 1982; Sætre *et al.* 1999; Veen *et al.* 2001; Svedin *et al.* 2008; Wiley *et al.* 2009; Ålund *et al.* 2013). Evolution of such strong intrinsic postzygotic isolation despite the recent divergence time between collared flycatcher and pied flycatcher makes these species unusual because diverging avian lineages are thought to develop intrinsic reproductive incompatibility more slowly (Price & Bouvier 2002; Fitzpatrick 2004). Therefore, rarity of backcross hybrids, coupled with the existence of strong postzygotic reproductive isolation, highlights that speciation progressed very rapidly in collared flycatcher and pied flycatcher.

The genetic basis for reduced fitness of hybrids could take other forms of incompatibility than a standard Bateson–Dobzhansky–Müller model where interacting loci are fixed for different alleles in hybridizing species. One such scenario is the case when one of the alleles (a_1), but not the other (a_2), at a polymorphic locus in one of the parental species shows reduced compatibility when interacting in a hybrid with a locus that is fixed for a species-specific allele (b_2) in the other parental species. If this were the case, we might expect to see a distortion in the segregation of a_1 and a_2 alleles when transmitted to hybrids. We tested for this scenario by comparing allele frequency distributions at segregating sites in collared flycatcher with the frequency distribution in hybrids of alleles transmitted by this species (Fig. 6). The test was limited to 23 846 loci where pied flycatchers were fixed for one of the alleles (a_2), meaning that the allelic contribution of the collared flycatcher parent could always be inferred. Four loci showed a highly significant deviation from the expected transmission ratio (Table 3). However, all of these four loci showed a deficit of the a_2 allele (i.e. an excess of the a_1 allele), which is not easily conceived under a model of incompatibilities between a_1 and variants of other loci transmitted by the pied flycatcher parent. We still think this is worthy of further investigation as the signal of biased transmission in each of the cases was very strong. Deviation from random inheritance of gametes can result from a number of mechanisms, including meiotic drive (de Villena & Sapienza 2001; Zollner *et al.* 2004; Huang *et al.* 2013).

Acknowledgements

This work was supported by the European Research Council (Grant number 249976), the Knut and Alice Wallenberg Foundation (Wallenberg Scholar Grant) and the Swedish Research Council (Grant number 2010-5650). Tomas Axelsson and Torbjörn Öst at the SNP&SEQ Technology Platform of Uppsala University are gratefully acknowledged for technical help.

References

Alatalo RV, Gustafsson L, Lundberg A (1982) Hybridization and breeding success of collared and pied flycatchers on the island of Gotland. *The Auk*, **99**, 285–291.

Alatalo RV, Eriksson D, Gustafsson L, Lundberg A (1990) Hybridization between pied and collared flycatchers – sexual selection and speciation theory. *Journal of Evolutionary Biology*, **3**, 375–389.

Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**, 1655–1664.

Alhaddad H, Khan R, Grahn RA *et al.* (2013) Extent of linkage disequilibrium in the domestic cat, *Felis silvestris catus*, and its breeds. *PLoS One*, **8**, e53537.

Ålund M, Immler S, Rice AM, Qvarnström A (2013) Low fertility of wild hybrid male flycatchers despite recent divergence. *Biology Letters*, **9**, 1–4.

Backström N, Ovarnström A, Gustafsson L, Ellegren H (2006) Levels of linkage disequilibrium in a wild bird population. *Biology Letters*, **2**, 435–438.

Backström N, Forstmeier W, Schielzeth H *et al.* (2010) The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Research*, **20**, 485–495.

Backström N, Sætre G-P, Ellegren H (2013) Inferring the demographic history of European *Ficedula* flycatcher populations. *BMC Evolutionary Biology*, **13**, 2.

de Bakker PIW, Yelensky R, Pe'er I *et al.* (2005) Efficiency and power in genetic association studies. *Nature Genetics*, **37**, 1217–1223.

Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.

Barton NH (2000) Genetic hitchhiking. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, **355**, 1553–1562.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 289–300.

Borge T, Lindroos K, Nadvornik P, Syvänen AC, Sætre GP (2005) Amount of introgression in flycatcher hybrid zones reflects regional differences in pre and post-zygotic barriers to gene exchange. *Journal of Evolutionary Biology*, **18**, 1416–1424.

Bourret V, Kent MP, Primmer CR *et al.* (2013) SNP-array reveals genome-wide patterns of geographical and potential adaptive divergence across the natural range of Atlantic salmon (*Salmo salar*). *Molecular Ecology*, **22**, 532–551.

Cantarel BL, Korf I, Robb SMC *et al.* (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, **18**, 188–196.

Davey JW, Hohenlohe PA, Etter PD *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, **12**, 499–510.

DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51–63.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756–760.

Fitzpatrick BM (2004) Rates of evolution of hybrid inviability in birds and mammals. *Evolution*, **58**, 1865–1870.

Fitzpatrick BM (2012) Estimating ancestry and heterozygosity of hybrids using molecular markers. *BMC Evolutionary Biology*, **12**, 131.

Gavrilets S (2004) *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton, New Jersey.

Gelter HP, Tegelström H (1992) High-frequency of extra-pair paternity in Swedish pied flycatchers revealed by allozyme electrophoresis and DNA fingerprinting. *Behavioral Ecology and Sociobiology*, **31**, 1–7.

Gompert Z, Lucas LK, Nice CC *et al.* (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Groenen MAM, Wahlberg P, Foglio M *et al.* (2009) A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Research*, **19**, 510–519.

Gunderson KL (2009) Whole-genome genotyping on bead arrays. *Methods in Molecular Biology*, **529**, 197–213.

Hagen IJ, Billing AM, Ronning B *et al.* (2013) The easy road to genome-wide medium density SNP screening in a non-model species: development and application of a 10K SNP-chip for the house sparrow (*Passer domesticus*). *Molecular Ecology Resources*, **13**, 429–439.

Harris RS (2007) *Improved Pairwise Alignment of Genomic DNA*. PhD Thesis, Pennsylvania State University, PA, USA.

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, **38**, 226–231.

Hill WG, Weir BS (1988) Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, **33**, 54–78.

Hohenlohe PA, Bassham S, Currey M, Cresko WA (2012) Extensive linkage disequilibrium and parallel adaptive divergence across threespine stickleback genomes. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **367**, 395–408.

Huang LO, Labbe A, Infante-Rivard C (2013) Transmission ratio distortion: review of concept and implications for genetic association studies. *Human Genetics*, **132**, 245–263.

Johnson GCL, Esposito L, Barratt BJ *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nature Genetics*, **29**, 233–237.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Karlsson S, Moen T, Lien S, Glover KA, Hindar K (2011) Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. *Molecular Ecology Resources*, **11**, 247–253.

Keller I, Wagner CE, Greuter L *et al.* (2013) Population genomic signatures of divergent adaptation, gene flow and hybrid speciation in the rapid radiation of Lake Victoria cichlid fishes. *Molecular Ecology*, **22**, 2848–2863.

Lundberg A, Alatalo RV (2010) *The Pied Flycatcher*. T & AD Poyser, London.

Nadachowska-Brzyska K, Burri R, Olason PI *et al.* (2013) Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genetics*, **9**, e1003942.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Parchman TL, Gompert Z, Braun MJ *et al.* (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, **22**, 3304–3317.

Price TD, Bouvier MM (2002) The evolution of F-1 postzygotic incompatibilities in birds. *Evolution*, **56**, 2083–2089.

Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: models and data. *American Journal of Human Genetics*, **69**, 1–14.

Purcell S, Neale B, Todd-Brown K *et al.* (2007) PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Qvarnström A, Bailey RI (2009) Speciation through evolution of sex-linked genes. *Heredity*, **102**, 4–15.

Qvarnström A, Rice AM, Ellegren H (2010) Speciation in *Ficedula* flycatchers. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **365**, 1841–1852.

Rundle HD, Nosil P (2005) Ecological speciation. *Ecology Letters*, **8**, 336–352.

Sætre GP, Sæther SA (2010) Ecology and genetics of speciation in Ficedula flycatchers. *Molecular Ecology*, **19**, 1091–1106.

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

Sætre GP, Kral M, Bures S, Ims RA (1999) Dynamics of a clinal hybrid zone and a comparison with island hybrid zones of flycatchers (*Ficedula hypoleuca* and *F. albicollis*). *Journal of Zoology*, **247**, 53–64.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual*, 2nd edn. Cold Spring Harbour Laboratory Press, New York.

Sheldon BC, Ellegren H (1999) Sexual selection resulting from extra-pair paternity in collared flycatchers. *Animal Behaviour*, **57**, 285–298.

Sherman JD, Stack SM (1995) Two-dimensional spreads of synaptonemal complexes from solanaceous plants. VI. High-resolution recombination nodule map for tomato (*Lycopersicon esculentum*). *Genetics*, **141**, 683–708.

Slatkin M (2008) Linkage disequilibrium – understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, **9**, 477–485.

Stapley J, Birkhead TR, Burke T, Slate J (2008) A linkage map of the zebra finch *Taeniopygia guttata* provides new insights into avian genome evolution. *Genetics*, **179**, 651–667.

Stumpf MPH, McVean GAT (2003) Estimating recombination rates from population-genetic data. *Nature Reviews Genetics*, **4**, 959–968.

Svedin N, Wiley C, Veen T, Gustafsson L, Qvarnström A (2008) Natural and sexual selection against hybrid flycatchers. *Proceedings of the Royal Society B-Biological Sciences*, **275**, 735–744.

Tortereau F, Servin B, Frantz L *et al.* (2012) A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, **13**, 586.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic Islands of speciation in Anopheles gambiae. *PLoS Biology*, **3**, e285.

Van Bers NEM, Santure AW, Van Oers K *et al.* (2012) The design and cross-population application of a genome-wide SNP chip for the great tit *Parus major*. *Molecular Ecology Resources*, **12**, 753–770.

Varshney RK, Song C, Saxena RK *et al.* (2013) Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology*, **31**, 240–246.

Veen T, Borge T, Griffith SC *et al.* (2001) Hybridization and adaptive mate choice in flycatchers. *Nature*, **411**, 45–50.

de Villena FPM, Sapienza C (2001) Nonrandom segregation during meiosis: the unfairness of females. *Mammalian Genome*, **12**, 331–339.

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L (2002) Distribution of recombination crossovers and the origin of haplotype blocks: the interplay of population history, recombination, and mutation. *American Journal of Human Genetics*, **71**, 1227–1234.

Wang L, Luzynski K, Pool JE *et al.* (2011) Measures of linkage disequilibrium among neighbouring SNPs indicate asymmetries across the house mouse hybrid zone. *Molecular Ecology*, **20**, 2985–3000.

Wiley C, Fogelberg N, Sæther SA *et al.* (2007) Direct benefits and costs for hybridizing *Ficedula* flycatchers. *Journal of Evolutionary Biology*, **20**, 854–864.

Wiley C, Qvarnstrom A, Andersson G, Borge T, Saetre GP (2009) Post-zygotic isolation over multiple generations of hybrid descendents in a natural hybrid zone: how well do single-generation estimates reflect reproductive isolation? *Evolution*, **63**, 1731–1739.

Zhao S, Zheng P, Dong S *et al.* (2013) Whole-genome sequencing of giant pandas provides insights into demographic history and local adaptation. *Nature Genetics*, **45**, 67–71.

Zollner S, Wen XQ, Hanchard NA *et al.* (2004) Evidence for extensive transmission distortion in the human genome. *American Journal of Human Genetics*, **74**, 62–72.

T.K. processed and analysed the genotype data, N.B. performed LD analyses, R.B. contributed to population genetic analysis, P.O. identified and selected SNPs for array design, A.H, A.M.R. M.Å. and A.Q. provided the samples and identified hybrids, T.K. and H.E. wrote the paper, with input from the other authors. H.E. conceived and designed the study.

Data Accessibility

The markers used in the array have been deposited to dbSNP under the accession numbers NCBI ss_974710772 -974751664. Individual genotypes have been deposited in the Dryad database (doi:10.5061/dryad.h68jd).

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1 Summary of analysis of tag SNPs for each chromosome and for all chromosomes combined (Total). No. SNPs = total number of SNPs genotyped in the population sample for each respective chromosome with minor allele frequency >10/%, No. tag SNPs = number of SNPs in the tagging panel at a particular tagging threshold (r2) that represent the entire set of SNPs for each respective chromosome (percentage of entire set within brackets).

Figure S1 Triangle plot indicating relationships between interspecific heterozygosity (H_I: vertical axis) and hybrid index [*S*: horisonal axis from 0 (pure collared flycatcher) to 1 (pure pied flycatcher)] of 48 putative hybrid individuals and five additional F_1-like hybrids identified in ‘pure’ parental species cohort by Admixture analysis.