FROM THE COVER
Genomic distribution and estimation of nucleotide diversity in natural populations: perspectives from the collared flycatcher (Ficedula albicollis) genome

LUDOVIC DUTOIT, RETO BURRI, ALEXANDER NATER, CARINA F. MUGAL and HANS ELLEGREN
Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden

Abstract
Properly estimating genetic diversity in populations of nonmodel species requires a basic understanding of how diversity is distributed across the genome and among individuals. To this end, we analysed whole-genome resequencing data from 20 collared flycatchers (genome size ≈1.1 Gb; 10.13 million single nucleotide polymorphisms detected). Genomewide nucleotide diversity was almost identical among individuals (mean = 0.00394, range = 0.00384–0.00401), but diversity levels varied extensively across the genome (95% confidence interval for 200-kb windows = 0.0013–0.0053). Diversity was related to selective constraint such that in comparison with intergenic DNA, diversity at fourfold degenerate sites was reduced to 85%, 3' UTRs to 82%, 5' UTRs to 70% and nondegenerate sites to 12%. There was a strong positive correlation between diversity and chromosome size, probably driven by a higher density of targets for selection on smaller chromosomes increasing the diversity-reducing effect of linked selection. Simulations exploring the ability of sequence data from a small number of genetic markers to capture the observed diversity clearly demonstrated that diversity estimation from finite sampling of such data is bound to be associated with large confidence intervals. Nevertheless, we show that precision in diversity estimation in large outbred population benefits from increasing the number of loci rather than the number of individuals. Simulations mimicking RAD sequencing showed that this approach gives accurate estimates of genomewide diversity. Based on the patterns of observed diversity and the performed simulations, we provide broad recommendations for how genetic diversity should be estimated in natural populations.

Keywords: genetic markers, nucleotide diversity, population genomics, recombination

Received 18 February 2016; revision received 2 September 2016; accepted 19 September 2016

Introduction
Genetic diversity is a key parameter in evolutionary biology and population genetics. It relates to the evolvability of populations (Fisher 1930), is important in the contexts of adaptation (Barrett & Schluter 2008), inference of population structure (Charlesworth 2010) and speciation (Coyne & Orr 2004), and is also relevant to conservation and management (Frankham 1995; Reed & Frankham 2003). Moreover, explaining the genetic diversity underlying phenotypic variation has long been a challenge to evolutionary biologists because directional as well as stabilizing selection should deplete this diversity (Barton & Turelli 1989; Barton & Keightley 2002). Knowledge about the levels and character of genetic diversity is important to questions like these, and consequently, the ability to accurately estimate genetic diversity is essential to the study of evolutionary phenomena.

In population genetic terms, genetic diversity reflects the interplay of mutation, genetic drift, selection, recombination and gene flow on DNA sequence variation. Intuitively, we expect large populations to harbour more genetic diversity than small populations and, in principle, this is defined by the population mutation rate (θ), which equals 4Neμ where Ne is the effective population size and μ is the rate of mutation. However, the determinants of genetic diversity are complex and it has long appeared mysterious that variation in levels of genetic diversity among species is relatively limited despite huge variation in population sizes (Leffler et al. 2012), an observation known as Lewontin’s paradox (Lewontin 1974). One possible explanation for this paradox is that the high genetic diversity expected for large populations is counteracted by genetic draft (Gillespie 2000, 2001), with selection being more efficient in large populations...
and thereby reinforcing the diversity-reducing effect of linked selection on neutral diversity (Corbett-Detig et al. 2015). Both pervasive positive selection (Maynard-Smith & Haigh 1974) and extensive purifying selection (Charlesworth 2012) will reduce neutral diversity at linked sites. Moreover, recent evidence suggests that life-history traits such as fecundity (Romiguier et al. 2014) rather than ecological disturbance and historical contingency (short-term variation in \( N_e \); (Banks et al. 2013)) may explain variation in diversity levels among species.

Besides variation in diversity levels among populations and species, it is important to note that a single \( \Theta \) cannot describe genomic diversity because selection locally reduces \( N_e \) to a different extent in different parts of the genome (Gossmann et al. 2011). Such effects are expected to be most pronounced in regions of low recombination where linked selection will affect the frequency of neutral variants over longer physical distances than in regions of high recombination. This prediction is supported by observations in several organisms of a positive correlation between recombination rate and nucleotide diversity (reviewed in (Cutter & Payseur 2013)). Additional factors that may contribute to within-genome variation in diversity levels include introgression, which may not be uniformly distributed across the genome (Wu & Ting 2004), and local or regional variation in the rate of mutation (Hodgkinson & Eyre-Walker 2011). Moreover, even under an idealized scenario of constant selection and mutation, stochastic variation in the coalescence process for individual genomic regions will render the amount of diversity variable across the genome (Wakeley 2009). On top of this, individuals within at least small populations can vary in their overall degree of genetic diversity due to different levels of inbreeding (Bensch et al. 2006) and this may also apply when there is selfing or frequent introgression. Yet, the ability to obtain a single value of the diversity parameter is important when broadly considering the effectiveness of selection and drift and in among-population comparisons.

For the reasons mentioned above, estimating genetic diversity is inherently sensitive to the number of sampled loci (because of heterogeneity in diversity levels across the genome) and individuals (because of potential variation in inbreeding among individuals), and also to the type of sequence analysed (e.g. neutrally evolving loci vs. sequences subject to selection). Theoretical work (Pluzhnikov & Donnelly 1996; Felsenstein 2006) and simulations (Carling & Brumfield 2007) have been used to describe the expected distribution of diversity levels across the genome and to determine how well finite sampling of loci and individuals is able to capture the variation in genetic diversity. However, it remained unclear to what extent these studies captured biologically relevant scenarios in terms of how genetic diversity in distributed across the genome in natural populations. An alternative way of analysing the extent to which genomic diversity is reflected in finite sampling from a population is to first empirically determine the genomewide landscape of diversity and then simulate sampling from the observed distribution. As this requires vast amounts of genomic data, it has up until now not been a viable option. However, this is bound to change given current progress in genome sequencing and resequencing of diverse groups of organisms (Ellegren 2014).

At this point, the collared flycatcher (Ficedula albicollis) is one of the ‘ecological model organisms’ in which the genome has been mapped in most detail. Following the generation of a draft sequence assembly of the 1.1 Gb collared flycatcher genome (Ellegren et al. 2012), the construction of a high-density linkage map using a 50-k single nucleotide polymorphism (SNP) array (Kawakami et al. 2014a) allowed the development of a second-generation assembly version with unusually high assembly continuity and with scaffolds ordered and oriented along chromosomes (Kawakami et al. 2014b; Smeds et al. 2014, 2015). In addition, whole-genome resequencing data from multiple populations and from related species are available (Burri et al. 2015; Nater et al. 2015; Kardos et al. 2016). This system therefore serves as a most useful model and offers excellent opportunities for studies of the landscape of genetic diversity in a eukaryotic genome. Here, we use whole-genome resequencing data from a population of collared flycatchers to address the following questions: How is genetic diversity distributed across chromosomes and the genome? Is the distribution of diversity more heterogeneous than expected by chance? To what extent does genomewide diversity vary among individuals and among functional categories of sequences? Then, based on real data, we use simulations to examine how different sampling schemes would affect estimates of genetic diversity and how sampling schemes can be optimized to capture most of the variation in genetic diversity within populations.

**Material and methods**

**Sequence analysis**

We used whole-genome resequencing data from 20 (10 males, 10 females) collared flycatchers from a single population in the Apennine mountain range in central Italy (available on the European Nucleotide Archive (ENA) under Accession Number PRJEB7359). Detailed information about sampling, DNA analyses and bioinformatic methods used for generating these data is given in Burri et al. (2015). Briefly, birds were sequenced with paired-end technology on an Illumina HiSeq 2000 instrument. Reads (2 × 100 bp, from libraries with insert sizes of
approximately 450 bp) were then mapped to a repeat-masked version of the collared flycatcher genome assembly FicAlb1.5 (GenBank Accession GCA_000247815.2) using BWA 0.7.5a (Li & Durbin 2009) and local realignment with GATK 3.2.2 (McKenna et al. 2010; DePristo et al. 2011). Variant calling was performed using a combination of UnifiedGenotyper in GATK, Samtools (Li et al. 2009), and FreeBayes (Garrison & Marth 2012) with default settings. Base quality score recalibration was then applied in two rounds. In addition to the procedures for variant calling described in (Burri et al. 2015), we applied a final filtering step by only including sites at which every individual was covered by at least five reads (with the exception of when comparing different individuals within windows).

Annotation

We downloaded ENSEMBL annotations for version FicAlb1.4 of the collared flycatcher genome assembly, translated them to the FicAlb1.5 assembly version and then categorized sequences as either representing intergenic regions, introns, coding sequences (CDS), or 5' or 3' untranscribed regions (UTR). Alternative transcripts of the same gene were concatenated. Within coding sequences, fourfold and 0-fold degenerate sites were distinguished. In doing so, codons with coding sequence on both strands and where more than one site was variable were removed from the data set.

Estimation of nucleotide diversity

As an estimator of the population mutation rate (θ), we estimated nucleotide diversity (π), the average pairwise number of differences per site among the chromosomes in a population (Nei & Li 1979).

\[
\pi = \sum_{i \neq j} x_i x_j \pi_{ij}
\]

where \( i, j \in \{1, \ldots, n\} \), and \( n \) is the number of sequences in the sample. Further, \( x_i \) and \( x_j \) denote the respective frequencies of the \( i \)th and \( j \)th sequences; \( \pi_{ij} \) denotes the number of nucleotide differences per nucleotide site between the \( i \)th and \( j \)th sequences.

We estimated nucleotide diversity along the genome in 200-kb nonoverlapping windows for every individual and for the population as a whole using in-house scripts and the Python package rvVCF 0.4.0 (https://pyvcf.readthedocs.org). Windows with <10 000 sites remaining after filtering were discarded. The chosen window size was a trade-off between capturing fine-scale variation in genetic diversity and reducing measurement error. In addition, separate estimates of genomewide nucleotide diversity were made for concatenated sequences of each of the sequence categories described above. Reported estimates of nucleotide diversity are at the population level. Estimates of individual nucleotide diversities are specified as such. For estimating Z chromosome diversity, we only used data from the 10 males; in birds, males are the homogametic sex.

To obtain a null expectation for the variation in genetic diversity along the autosomal genome, we randomly reassigned the location of the total number of observed SNPs. We performed 10 resamplings and extracted the distribution of nucleotide diversity within windows under this random distribution of segregating sites for every replicate.

To examine whether any of the individuals were closely related, we assessed relatedness among individuals by calculating of the unadjusted Ajk statistic (Yang et al. 2010) and inbreeding was assessed by \( F_{IS} \) using the Hierfstat (Goudet 2005) package for R (R Development Core Team 2014).

Simulation of diversity estimation

To simulate the precision of nucleotide diversity estimates from a given number of markers and individuals, we subsampled the empirical data under different sampling schemes designed to correspond to amplicon sequencing and RAD sequencing, respectively.

In the schemes mimicking amplicon sequencing, we defined a marker as 500 sites for which all the sampled individuals had a coverage at least five reads in a physical region of maximum 1500 bp. Sampling schemes consisted of 1–20 individuals sampled for 1–50 ‘markers’ (in this case regions of the genome), done separately for coding sequences, intergenic regions, introns and 5'UTRs, and 1, 5, 10 and 20 individuals for 50–200 markers in intergenic regions. The schemes mimicking RAD sequencing consisted of 10 or 20 individuals sampled for 500–5000 100-bp markers of random sequence for which each bp had a coverage of at least five reads for all the sampled individuals.

For all sampling schemes, subsampling was repeated 1000 times and nucleotide diversity estimated for each sampled set of individuals and markers. The interval covering 95% of the distribution of nucleotide diversity estimates was recorded. Previous studies examining the variance in estimates of nucleotide diversity from the use of a finite number of markers have referred to this as accuracy of precision in diversity estimation (Pluzhnikov & Donnelly 1996; Felsenstein 2006; Carling & Brumfield 2007). Here, we refer to the width of the 95% confidence intervals (CIs) as precision in diversity estimation as it is a measure of the consistency of \( \pi \) across subsamples for a given sampling scheme.
Results and discussion

Overall levels of genetic diversity

Whole-genome resequencing of 20 collared flycatchers resulted in a mean per-site coverage of 13.9 x across individuals, with a range of 7.6–21.4 x. 575.64 Mb out of the 1.047 Gb total autosomal assembly met the criteria of having a read coverage of at least 5 x in all 20 individuals, and 10.13 million of these sites were variable in our sample. The large number of SNPs provides an excellent opportunity to assess genomewide diversity at the population as well as at the individual level. None of the individuals were closely related (maximum pairwise AJK value: 0.09 Fig. S1, Supporting information), and there was limited evidence for inbreeding (\( F_{IS} = 0.0037 \)). This suggests that the analysed individuals constitute a random sample of birds from the population. Estimates of genomewide, autosomal nucleotide diversity per individual were almost identical among the 20 individuals (mean \( \pi = 0.00386 \), range = 0.00376–0.00393; Fig. 1A, Table S1, Supporting information) and corresponded to a mean of one heterozygous position every 259 bp in an individual’s genome (Table 1). With all 20 individuals taken together, the mean density of segregating sites in the sample was one every 103 bp. A survey of diversity data from 167 species representing 14 phyla found that the majority of species have \( \pi \) in the range of 0.0005–0.05, with a median of \( \approx 0.0065 \) (Leffler et al. 2012). Collared flycatcher thus shows intermediate levels of genetic diversity seen in this broader context. As expected from the lower effective population size, nucleotide diversity along the Z chromosome was lower than on autosomes (mean \( \pi = 0.00288 \)). As the evolutionary forces affecting diversity levels of autosomal and sex-linked sequences differ, and addressing these differences was beyond the scope of the study, we will not further deal with sex-linked sequences.

Genetic diversity in different sequence categories

Selection will generally reduce diversity levels. Overall, nucleotide diversity was strongly dependent on the annotated type of sequence and was directly related to the expected selective pressure (Table 2), with coding sequences being the least variable category (mean \( \pi = 0.0014 \)). Using intergenic sequence as a putatively neutral reference (mean \( \pi = 0.0040 \)), diversity at fourfold degenerate sites was reduced to 85% (0.0034), 3' UTRs to 82% (0.0033), 5' UTRs to 70% (0.0028) and nondegenerate sites to 12% (0.0005). Diversity in introns (mean \( \pi = 0.0039 \)) was similar to that in intergenic regions. As these two categories of sequences are by far most common in avian genomes, they largely determined the genomewide average.

The lower diversity at fourfold degenerate (silent) sites compared to intergenic and intronic sites may seem surprising at first given that all three sequence categories traditionally have been considered to evolve neutrally. However, it supports observations in molecular evolutionary studies of birds (Künstner et al. 2011) and several other organisms (Eöry et al. 2010; Pollard et al. 2010;
Table 1  Average (±SD) number of variable sites per individual for each type of annotated sequence category extrapolated for the whole genome

<table>
<thead>
<tr>
<th>Sequence category</th>
<th>Number of variable sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′ UTR</td>
<td>15 698 ± 225</td>
</tr>
<tr>
<td>CDS</td>
<td>14 343 ± 193</td>
</tr>
<tr>
<td>3′ UTR</td>
<td>865 ± 29</td>
</tr>
<tr>
<td>Introns</td>
<td>754 081 ± 8020</td>
</tr>
<tr>
<td>Intergenic regions</td>
<td>1 482 708 ± 17 170</td>
</tr>
<tr>
<td>Total</td>
<td>2 267 696 ± 25 208</td>
</tr>
</tbody>
</table>

Table 2  Nucleotide diversity (π) in annotated regions of the collared flycatcher genome, and their fraction of the whole genome

<table>
<thead>
<tr>
<th>Sequence category</th>
<th>π</th>
<th>Proportion of genome (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intergenic</td>
<td>0.0040</td>
<td>65.9</td>
</tr>
<tr>
<td>Intrinsic</td>
<td>0.0039</td>
<td>30.7</td>
</tr>
<tr>
<td>Fourfold degenerated sites</td>
<td>0.0034</td>
<td>1.3</td>
</tr>
<tr>
<td>Nondegenerated (0-fold) sites</td>
<td>0.0005</td>
<td>0.4</td>
</tr>
<tr>
<td>5′ UTR</td>
<td>0.0026</td>
<td>1.0</td>
</tr>
<tr>
<td>3′ UTR</td>
<td>0.0033</td>
<td>0.1</td>
</tr>
<tr>
<td>Total</td>
<td>0.0039</td>
<td></td>
</tr>
</tbody>
</table>

Data are for autosomes.

Clemente & Vogl 2012; Lawrie et al. 2013) indicating that silent sites are constrained by purifying selection. For example, they might be functionally involved with splicing or mRNA stability (Chamary et al. 2006). An alternative but not mutually exclusive explanation is that close linkage to nearby nonsynonymous sites under selection might reduce diversity at synonymous sites due to hitchhiking effects (Cutter & Payseur 2013; Corbett-Detig et al. 2015). However, if linked selection is pervasive, this should also to some extent affect introns.

It is noteworthy that regulatory sequences subject to purifying selection reside within introns and intergenic regions (Ponting 2008), illustrating the difficulty in defining a truly neutral category of sequences. One way of handling this in attempts to estimate the neutral substitution rate (and thereby getting a proxy for the mutation rate) is to mask noncoding regions that are highly conserved in alignments to sequences from distantly related species (Siepel et al. 2005). The downside of this approach is that it is difficult to distinguish between weak selective constraints and locally reduced mutation rates as the cause of sequence conservation. Moreover, functional elements in noncoding DNA may be lineage-specific due to rapid turnover (Smith et al. 2004; Meader et al. 2010) and therefore not detectable as conserved regions in distant alignments.

Finally, we note that many molecular evolutionary studies, including in birds (Kunstner et al. 2011), have revealed evolutionary constraints in UTRs. This is consistent with the reduced diversity in these regions that we observe. Among other functions, UTRs are likely to contain regulatory sequences affecting gene expression including binding sites for micro-RNAs (Bartel 2009).

Variation in genetic diversity across the genome

A comparison of the simulated distribution of π estimates among nonoverlapping 200-kb windows assuming random location of segregating sites in the genome and the observed distribution of π estimates clearly showed that genetic diversity is heterogeneously distributed across the genome (Fig. 1B). This was also evident from a significant autocorrelation in π estimates between neighbouring 200-kb windows (Fig. 1C), demonstrating regional similarity in diversity levels. On average, there was a statistically significant correlation in π estimates between windows located up to 1 Mb apart (Fig. 1C). The similarity in diversity levels on a regional scale, but also the extensive variation on a larger scale, can readily be seen from the distribution of π estimates along a chromosome (Fig. 2A). Diversity estimates differed by nearly two orders of magnitude between the least and the most variable window, with a 95% CI of 0.0013–0.0053.

The variation in π among windows was not merely a consequence of different proportions of less variable coding sequences as excluding coding sequences from the estimate did not change the density distribution of nucleotide diversity (Fig. 1B). Moreover, as coding sequences only constitute 2% of the flycatcher genome, and UTRs another 1%, the proportion of coding sequence per window has limited influence on the variation in diversity among 200-kb windows (while it would have had higher influence at smaller window sizes). However, as shown below, the density of coding sequence may still be important in determining diversity levels via linked selection. Although window-based π estimates of different individuals were highly correlated (Fig. 2B; Spearman’s ρ, range = 0.70–0.86), there was far more variation among individuals at the level of windows than at the level of the whole genome (Fig. 2A). This may of course be a statistical effect of the much larger number of sites involved when the whole genome is considered but variation in coalescence times among haplotypes, for example due to selection, could add another layer of variation among individuals in different genomic regions. The mean difference between the highest and lowest individual π estimate per window was 0.0018 ± 0.0012.

To address the causes of the broad-scale variation in genetic diversity across the genome, we analysed the relationship between chromosome size and diversity. This analysis was motivated by the fact that chromosome
size and recombination rate are negatively correlated in collared flycatchers (Kawakami et al. 2014b) as well as in many other species of birds (e.g. ICGSC 2004) and other organisms (e.g. Farré et al. 2012; Jensen-Seaman et al. 2004). In flycatchers, the variation in mean recombination rate per chromosome is extensive, from ≈ 2 cM/Mb for chromosomes >100 Mb to >10 cM/Mb for chromosomes <10 Mb (Kawakami et al. 2014b). High rates of recombination in small chromosomes should be associated with high levels of genetic diversity (i.e. there should be a negative correlation between genetic diversity and chromosome size) because recombination uncouples selected and neutral loci, thereby reducing the effect of linked selection on neutral diversity. Studies in many species of animals and plants (reviewed in Cutter & Payseur 2013), including flycatchers (Burri et al. 2015), have indeed demonstrated a positive correlation between rate of recombination and genetic diversity. However, we found a positive relationship between chromosome size and π, with diversity increasing from a chromosome average of π ≈ 0.0025 for the smallest chromosomes (<5 Mb) to reach a plateau at ≈ 0.0040–0.0043 for chromosomes in the size range of 50–150 Mb (Fig. 3A; Spearman’s ρ = 0.803, P < 0.0001). A positive correlation was seen for all sequence categories (coding sequences, fourfold degenerate sites, nondegenerate sites, 5’UTR, 3’UTR, introns, intergenic DNA) when analysed separately (Fig. S3, Supporting information).

![Fig. 2](A) Example of variation in nucleotide diversity of a chromosomal segment shown for 200-kb windows within position 30–60 Mb of chromosome 1. The solid line represents the population median, and the grey lines delimit the maximum and minimum individual nucleotide diversity among 20 birds. (B) Example of the correlation between individual nucleotide diversity in 200-kb windows of two birds (Pearson’s r = 0.87).

![Fig. 3](A) Correlation between nucleotide diversity and chromosome size (log10 scale) in (A) collared flycatcher (this study), (B) brown creeper (Certhia americana) with data from (Manthey et al. 2015) and (C) Hawaii amakihi (Hemignathus virens) with data from Callicrate et al. (2014).
The positive relationship between chromosome size and nucleotide diversity indicates that factors other than recombination are related with chromosome size and affect diversity. One such factor is the density of targets for selection, which can be approximated by the density of coding sequence. There was a strong negative correlation between the density of coding sequence and chromosome size in flycatchers (Fig. 4), consistent with findings in other avian genomes [e.g. (ICGSC 2004)]. Everything else being equal, the higher the density of sequences subject to selection, the stronger the diversity-reducing effect of linked selection, which is observed in this system (Burri et al. 2015). The opposing effects of high rates of recombination and high density of targets of selection on levels of nucleotide diversity in small avian chromosomes suggest that the direction of the correlation between nucleotide diversity and chromosome size will be governed by the relative strength of the effects of these two causative factors. Without detailed investigation, it is hard to predict which of the effects would dominate and it might very well be that the dominating factor varies among avian species. For example, as the diversity-reducing role of linked selection should be more prevalent when \(N_e\) is large (Corbett-Detig et al. 2015), it might be that the importance of the density of targets for selection increases with increasing \(N_e\). In our study system, it seems that the effect from density of targets of selection prevails because we found a positive correlation between nucleotide diversity and chromosome size. Analyses using the pairwise sequentially Markovian coalescent (PSMC) indicate that the long-term \(N_e\) of the investigated collared flycatcher population has been \(\approx 500,000\) (Nadachowska-Brzyska et al. 2016).

To further investigate the relationship between chromosome size and diversity, we searched the literature for data on levels of nucleotide diversity per chromosome in other bird species. Figure 3B (Certhia americana (Manthey et al. 2015)) and C (Hemignathus virens (Callicrate et al. 2014)) show the relationship between nucleotide diversity and chromosome size for two species of the order Passeriformes. In Certhia americana, in contrast to flycatchers, the correlation is negative, while in Hemignathus virens there is no significant relationship. In chicken, one study found no significant relationship between SNP rate and chromosome size [figure 2 in (ICGPC 2004)], but using data from dbSNP (http://www.ncbi.nlm.nih.gov/snp/), we actually find a negative correlation between the number of SNPs per bp and chromosome size in chicken (Spearman’s rho = –0.723, \(P = 6 \times 10^{-9}\)). Using the same statistic for flycatchers (the number of SNPs per bp), we still find a strong positive correlation between diversity and chromosome size (rho = 0.905; \(P < 0.001\); Fig. S2, Supporting information).

Thus, the data available so far suggest that the relationship between nucleotide diversity and chromosome size in birds is inconsistent among species.

The allele frequency spectrum

With complete sequence data from 20 diploid individuals (40 chromosomes), we could obtain a detailed picture of the genomewide allele frequency spectrum, both overall and for specific sequence categories (Fig. 5). Awareness of the shape of the frequency spectrum is important when selecting informative SNP markers for inclusion in genotyping arrays, but can also inform about selection and demography. The spectrum was strongly shifted towards low frequency alleles with about 25% of all segregating sites in intergenic and intronic sequences being singletons (i.e. corresponding to a minor allele frequency of 0.025 in our sample of 40 chromosomes). As expected for sequences evolving under purifying selection, coding sequence alleles (and especially so for nondegenerate sites) segregated at lower frequencies than alleles in intergenic and intronic DNA (Fig. 5A). A slight shift towards low frequency alleles for fourfold degenerate sites (compared to intergenic and intronic DNA) suggests higher selective constraint on these sites, in line with the observation of reduced nucleotide diversity level at fourfold degenerate sites (Fig. 5B). Alternatively, close linkage to nonsynonymous sites under selection might affect the allele frequency spectrum also of silent sites.

Simulation of \(\pi\) estimation with genetic markers

Having characterized the landscape of genetic diversity in the flycatcher genome by whole-genome analysis, we set out to test how well analyses of a finite number of loci, such as often applied in ecological or evolutionary studies, would be able to capture the observed
genomewide diversity in the population. To mimic a situation of analysis of genetic diversity based on amplicons (PCR-amplified markers), we used different sampling schemes of 1–20 individuals sequenced for up to 200–500-bp regions randomly placed in the genome.

We refer to these regions as ‘markers’. We ran separate simulations for intergenic DNA, introns, coding sequences and UTRs, respectively.

An overall and indeed important observation from the simulation of sampling variance was that the 95% CI of \( \pi \) estimation was relatively broad for many sampling schemes, in particular when the number of markers was limited. This reflects that there is significant variation in diversity levels across the genome (Fig. 6A,B). Moreover, our results show that the precision in diversity estimation increased more by adding markers than by adding individuals (Figs 6 and 7), consistent with the observation of higher variation in diversity levels along the genome as compared to variation among individuals. For example, with 10 intergenic markers, the 95% CI for \( \pi \) estimation was similar for sequence data from only one individual (0.0018–0.0068; a ratio of 3.7 between upper and lower 2.5% tail) and from 10 individuals (0.0027–0.0055; 2.9). In contrast, with sequence data from one marker, the 95% CI for 10 individuals was very much wider (0.0004–0.009; 22.5) than that for 10 markers (0.0027–0.0055; 2.9). The precision in the estimation of \( \pi \) showed a steep increase (narrower CI) in the range of 1–5 markers, while further addition of markers only slowly led to improved precision (Fig. 6A,B). However, even with as much as 50 markers (i.e. a total of 25 kb) sequenced for 10 individuals, the ratio between the upper and lower 95% CI tails of \( \pi \) estimation was 1.3, and for 200 markers 1.2, giving an indication of how many markers are needed to accurately estimate genomewide diversity. The same trends were seen for all sequence categories (Fig. 7), but it should be noted that the precision in the estimation of \( \pi \) was higher for selectively constrained markers (in particular for coding sequences) than for intergenic and intronic DNA (Fig. 7; compare the size of CIs towards the lower left corner of the different plots). This is probably an effect of the lower genetic diversity of coding sequences leading to lower variance.

Simulations meant to mimic RAD sequencing (500–5000 markers) gave narrow confidence intervals of \( \pi \) estimates throughout the whole marker range (Fig. 6C). The ratio between the upper and lower CI tails of \( \pi \) estimation was only 1.2 for 500 markers (0.0035–0.0042) and 1.1 for 5000 markers (0.0037–0.00409). It is noteworthy that high precision was obtained despite only 100 bp were analysed per sampled marker, as is commonly the case in next-generation sequencing. Again this is consistent with the higher precision provided by the sampling of many markers. It could also be noted that as the RAD sequencing simulations sampled random sequences across the genome and thus included data from all sequence categories, \( \pi \) estimates converged towards somewhat lower values than was the case when amplicon sequencing of intergenic sequences was simulated (Fig. 6).
Recommendations for sampling design

Under the conditions of our study system (an outbred population with low levels of linkage disequilibrium; Backström et al. 2006) and simulation design, the recommendation for diversity estimation would be to put more emphasis on the number of loci analysed than on the number of individuals sequenced. This recommendation should be relatively insensitive to the precise length of individual markers used and we see no reason that it would not be broadly applicable across organisms and populations. Variation in \( N_e \) across the genome, beyond stochastic variation in the coalescence process, is likely to be a common feature of many organisms and generating sequence data from a large number of individuals from the same loci is essentially adding nonindependent data because they share a common evolutionary history (Pluzhnikov & Donnelly 1996). Moreover, we argue that significant heterogeneity in diversity levels across the genome implies that, given a fixed total amount of sequence obtained, increasing the amount of analysed sequence per locus does not generally make up for reducing the number of loci. The simulations mimicking RAD sequencing clearly showed that even with only 100 bp sequenced per loci, precision in \( \pi \) estimation was high when a large number of loci were sampled. In practice, our results demonstrate that RAD sequencing data will accurately reflect genomewide diversity and that little is
to be gained in this respect by whole-genome sequencing as far as mean levels of diversity are concerned.

Our recommendations concord with the conclusions from theoretical work by (Pluzhnikov & Donnelly 1996) and (Felsenstein 2006). (Carling & Brumfield 2007) used simulations to test the influence of the number and length of markers, and the number of individuals, on accuracy of estimates of genetic diversity. Overall, their work also supported the use of more markers over more individuals, but in contrast to our study, their simulations were based on simulated sequence data, not empirical data. The strong and heterogeneous effects of linked selection across the genome mean that it can be difficult to simulate sequence evolution in a way that the extent and character of polymorphism mirror biologically meaningful scenarios.

One situation at which the recommendation could potentially be compromised is for populations with significant inbreeding and, in particular, variation in the extent of inbreeding among individuals. In such cases, the importance of analysing more individuals in order to get a representative estimate of the genetic diversity in the population increases. Another aspect of inbred populations is the increased proportion of the genome that is identical by descent (IBD), meaning that the distribution of levels of genetic diversity within the genome starts to become detectably bimodal, with a certain fraction devoid of variability due to runs of homozygosity in IBD tracts (Pemberton et al. 2012). We also note that for several other purposes related to population genetic (e.g. for high resolution estimation of the allele frequency spectrum, and statistical tests based on such spectra) and/or molecular evolutionary analyses (e.g. like in the case of the McDonald–Kreitman test), it is more critical to sample a large number of individuals than in the case of estimating nucleotide diversity (Robinson et al. 2014).

Our simulations were based on sequence data from markers randomly distributed across the genome. With considerable variation in diversity levels among genomic regions and chromosomes, a nonrandom selection of markers for characterization of genetic diversity in a population could easily bias diversity estimates, or at least make them less comparable to estimates based on different sets of markers in other species. A correlation between chromosome size and genetic diversity seems particularly relevant to consider in this respect because the direction of correlation is apparently not consistent among species. For this reason, the recommendation would be to sample loci from as wide a range of chromosomes as possible.

Comparisons of diversity levels among species could also be hampered if the markers used are heterogeneous with respect to the type of annotated sequence, for example, if they contain both coding sequence and UTRs, or both coding and intronic sequence. If the goal is to estimate neutral genetic diversity, noncoding loci well distributed across the genome should therefore be targeted. This speaks in favour of using RAD sequencing or whole-genome resequencing for this purpose. It may be particularly relevant to emphasize that our results demonstrate that synonymous sites are generally less variable than intergenic regions and introns, indicating that they are evolving under the influence of purifying selection. Thus, such sites do not represent an ideal category of sequences for estimating neutral diversity. Finally, we stress that the analyses and simulations presented here, and the recommendations based on them, are for autosomal sequences. Many of the evolutionary forces affecting sequence diversity differ between sex chromosomes and autosomes.

Conclusions

We have shown that there is extensive variation in the level of nucleotide diversity across the genome of an avian species. This variation is seen in autosomal sequence and is thus unrelated to the well-known effects of sex linkage on genetic diversity (Hedrick 2007; Frankham 2012). Linked selection is likely to play a strong role in governing within-genome heterogeneity in diversity levels, with (variation in) recombination rate and density of targets of selection being primary determinants of the extent of linked selection. As far as we are aware, this study is the first to characterize genomewide nucleotide diversity through whole-genome resequencing of a large population sample and then use these real data to simulate how well genetic diversity would be captured by the use of genetic markers. We find that diversity estimation by sequencing a small number of amplicons is bound to be associated with large confidence intervals. Given the heterogeneity in diversity levels across the genome, gathering sequence data from many loci will increase the precision in diversity estimation. Naturally, one could ask whether molecular ecological studies will continue to be based on sequence data from a limited number of loci when genotyping-by-sequencing and whole-genome resequencing become increasingly feasible in many projects. However, even with the use of next-generation sequencing technologies, target capture approaches are cost-effective and can be used for a wide range of applications (Jones & Good 2016).

The ability to reliably estimate genetic diversity of different populations is critical for making conclusions about evolutionary processes. To end with an example (Gohli et al. 2013), recently reported an association between genetic diversity and female promiscuity in 18 passerine bird species based on sequence data from five introns (mean length $\approx 400$ bp). One possible explanation
to this would be that species with strong sexual selection for compatible genes (i.e. negative frequency-dependent selection for rare or dissimilar alleles) have relatively high levels of genetic diversity. The validity of this result was criticized by (Spurgin 2013) on several methodological grounds, including the precision of diversity estimates and the inference of species level diversity from sampling of individual populations (see also response to the criticism by (Lifjeld et al. 2013). Based on experiences from the present study, we note that more firm conclusions should have been possible to reach with more extensive sampling of genomic data, either confirming or disproving the idea of a relationship between genetic diversity and female promiscuity.

Acknowledgements

This study was funded by the Swedish Research Council (grant numbers 2010-565 and 2013-8271) and the Knut and Alice Wallenberg Foundation (KAW Scholar grant). Sequencing was performed by the SNP&SEQ Technology Platform at Uppsala University, as part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. Computations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) through Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX). We thank Rory Craig for comments on the manuscript.

References


© 2016 The Authors. Molecular Ecology Resources Published by John Wiley & Sons Ltd.
L.D. performed all population genomic analyses. R.B. provided access to data, A.N. and C.F.M. provided input on study design, data interpretation and statistical analyses. H.E. conceived of the study and supervised the work. L.D. and H.E. initiated the study and wrote the paper.

Data accessibility
Sequences are available in the European Nucleotide Archive (ENA), Accession Number PRJEB7359. Scripts and details of the data are made available at Dryad: doi:10.5061/dryad.1n84v.

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

Table S1 Genome-wide nucleotide diversity per individual

Fig. S1 Pairwise relatedness among individuals in the studied collared flycatcher population

Fig. S2 Number of SNPs per bp in relation to chromosome size in the collared flycatcher genome

Fig. S3 Relationship between chromosome size (log_{10}) and nucleotide diversity for different sequence categories. All correlations were significant (P < 0.013, or less). The strength of correlations were: coding sequences (R^2 = 0.68), 0-fold degenerate sites (R^2 = 0.62), fourfold degenerate sites (R^2 = 0.69), 5' UTR sites (R^2 = 0.44), 3' UTR sites (R^2 = 0.80) intergenic DNA (R^2 = 0.18) and introns (R^2 = 0.29)

© 2016 The Authors. Molecular Ecology Resources Published by John Wiley & Sons Ltd.