PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Dutoit L, Vijay N, Mugal CF, Bossu CM, Burri R, Wolf J, Ellegren H. 2017 Covariation in levels of nucleotide diversity in homologous regions of the avian genome long after completion of lineage sorting. *Proc. R. Soc. B* **284**: 20162756. http://dx.doi.org/10.1098/rspb.2016.2756

Received: 12 December 2016
Accepted: 18 January 2017

Subject Category:
Genetics and genomics

Subject Areas:
evolution, genetics, genomics

Keywords:
nucleotide diversity, linked selection, recombination rate, birds

Author for correspondence:
Hans Ellegren
e-mil: hans.ellegren@ebc.uu.se

†Present address: Lab of Molecular and Genomic Evolution, Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA.
‡Present address: Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden.
§Present address: Institute of Ecology, Department of Ecology, Friedrich Schiller University, Jena, Dornburger Strasse 159. 07743 Jena, Germany.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3677089.

THE ROYAL SOCIETY PUBLISHING

Covariation in levels of nucleotide diversity in homologous regions of the avian genome long after completion of lineage sorting

Ludovic Dutoit[1], Nagarjun Vijay[1,†], Carina F. Mugal[1], Christen M. Bossu[1,‡], Reto Burri[1,§], Jochen Wolf[1,2] and Hans Ellegren[1]

[1]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36 Uppsala, Sweden
[2]Division of Evolutionary Biology, Faculty of Biology II, Ludwig-Maximilians-Universität München, Grosshaderner Strasse 2, 82152 Planegg-Martinsried, Germany

HE, 0000-0002-5035-1736

Closely related species may show similar levels of genetic diversity in homologous regions of the genome owing to shared ancestral variation still segregating in the extant species. However, after completion of lineage sorting, such covariation is not necessarily expected. On the other hand, if the processes that govern genetic diversity are conserved, diversity may potentially covary even among distantly related species. We mapped regions of conserved synteny between the genomes of two divergent bird species—collared flycatcher and hooded crow—and identified more than 600 Mb of homologous regions (66% of the genome). From analyses of whole-genome resequencing data in large population samples of both species we found nucleotide diversity in 200 kb windows to be well correlated (Spearman's $\rho = 0.407$). The correlation remained highly similar after excluding coding sequences. To explain this covariation, we suggest that a stable avian karyotype and a conserved landscape of recombination rate variation render the diversity-reducing effects of linked selection similar in divergent bird lineages. Principal component regression analysis of several potential explanatory variables driving heterogeneity in flycatcher diversity levels revealed the strongest effects from recombination rate variation and density of coding sequence targets for selection, consistent with linked selection. It is also possible that a stable karyotype is associated with a conserved genomic mutation environment contributing to covariation in diversity levels between lineages. Our observations imply that genetic diversity is to some extent predictable.

1. Introduction

Understanding the evolutionary mechanisms governing the extent of genetic diversity (e.g. degree of polymorphism, heterozygosity or nucleotide diversity) within and between species is important to evolutionary biology in several respects [1]. For example, genomic scans for adaptively evolving loci require distinguishing signals of selection from other factors influencing genetic diversity [2]. Studies of population differentiation and speciation genetics are based on patterns of diversity and divergence, and the relationship between these parameters, such as the estimation of F_{ST} [3]. Moreover, genetic diversity is essential to conservation biology, including questions related to inbreeding and to the long-term adaptability of endangered species [4].

Genetic diversity is not a constant entity across the genome, but is known to vary considerably among chromosomes, genomic regions and functional categories of sequences [5–8]. As long as ancestral variation still segregates in

diverging lineages (i.e. lineage sorting is not completed), levels of genetic diversity in homologous regions of diverging genomes might be correlated. However, once ancestral variation is no longer shared owing to fixation of previously segregating variants, there is no reason *a priori* to expect diversity levels of homologous regions to covary between species. Yet, if the patterns and processes that govern diversity levels within genomes are conserved over evolutionary time scales, then diversity levels might be correlated. One notable situation concerns orthologous sites and sequences evolving under purifying selection in parallel lineages—such sequences are expected to show reduced nucleotide diversity in both lineages. Genes and other functional elements common to species are examples of sequences that are likely to show similarly low levels of diversity in different lineages. On the other end of the diversity spectrum, some sites and sequences could show increased diversity in parallel lineages owing to balancing selection [9]. However, trans-species polymorphisms would have to represent a high proportion of all polymorphisms to cause covariation of genetic diversity between species.

Even for neutrally evolving sequences genetic diversity of homologous regions of diverging genomes could potentially covary. One possible reason for this would be the local mutation rate (μ; cf. $\theta = 4\,N_e\mu$), which varies across the genome [10,11]. Another factor would be the degree of linked selection [12], which reduces diversity levels through background selection [13] or selective sweeps [14,15]. If the patterns of mutation rate variation and/or the intensity of linked selection are conserved among species, then this might result in covariation in neutral diversity levels.

Compared with other vertebrates, avian genomes are recognized to have unusually stable karyotypes [16] with $2n$ (diploid number of chromosomes) = 76–80 in the majority of species [17]. Essentially, all species characteristically show a limited number of large chromosomes (macrochromosomes) and a large number of very small chromosomes (microchromosomes) [18]. If karyotypic stability is associated with conservation of evolutionary processes governing genetic diversity (see further below), we hypothesized that covariation in regional levels of genetic diversity might be detectable in diverging lineages of birds. Here, we test this hypothesis using whole-genome resequencing data from population samples of two distantly related passerine species, the collared flycatcher (*Ficedula albicollis*) and the hooded crow (*Corvus (corone) cornix*), both species with a karyotype of $2n = 80$ [19]; for flycatcher karyotype information is from *F. parva* and *F. mugimaki* [20]. Genome assemblies with high sequence continuity are available for both species [21–23], and both genomes have been functionally annotated [24]. Phylogenetic analyses place the separation between the two lineages in the order of 25 million years ago (Ma) [25–27], which should be seen as a minimum time of divergence, because fossils put the early core corvids at 20–25 Ma [27]. Crow–flycatcher divergence thus corresponds to at least 4–12 million generations assuming a generation time of 6 years for hooded crows [28] and 2 years for flycatchers [29]. With an estimated long-term N_e of 200 000 for both species [30–32], this yields a range of 20–60 N_e generations as time to the most common ancestor. Because this is clearly beyond the expected time for complete lineage sorting (9–12 N_e generations [33]), the two species are thus not expected to share neutral ancestral polymorphism.

2. Material and methods

(a) Identification of genomic regions of conserved synteny

We identified regions that shared the same ancestral localization between the hooded crow (assembly v. 2.7) and collared flycatcher (fAlb15). We referred to these as regions of conserved synteny, and did not proceed to a base-to-base alignment as the synteny approach is much simpler and sufficient for the question addressed in the study.

First, we obtained pairwise alignments using LastZ v. 1.02.00 [34] and repeat-masked genome assemblies. We then used the UCSC Genome Browser toolset [35] and the JCVI library [36] in order to obtain a chain file, an alignment that allows gaps in both sequences at the same positions. This chain file was then hierarchically reorganized to be used as a lift-over chain (i.e. the conversion file to translate genomic coordinates from one species to the other according to conserved synteny between the two genomes).

We used liftOver, a program from the UCSC Kent source utilities package [35], to convert regions from one genome into the other. We used non-overlapping 200 kb windows along autosomes of the flycatcher genome as reference and retrieved conserved syntenic, collinear sequences in the crow. Windows were retained for further analyses if more than 80% of the bases in collared flycatcher could be remapped to one window in the crow. We excluded alignments less than 180 kb or greater than 220 kb as large size discrepancies may not only indicate repeat region reductions/expansions or small rearrangements, but could also be a sign of spurious alignments.

(b) Population re-sequencing data

We extracted re-sequencing data from 30 hooded crows sampled in two populations (Poland and Sweden) and 30 collared flycatchers also sampled in two populations (Czech Republic and Italy). Procedures for read mapping and variant calling are described in the original reports of polymorphism data [23,37] and were largely consistent between the two datasets. Briefly, raw reads were mapped to the respective reference genomes using BWA [38] v. 0.7.4 followed by local realignment using GATK [39,40] (v. 2.3.6 for hooded crows and v. 2.4.9 for collared flycatchers) and removal of duplicates using PICARD (http://picard.sourceforge.net), v. 1.46 for crows and v. 1.77 for flycatchers. Variant discovery was performed on a per-population basis to account for population structure. For both species, base quality score recalibration (BQSR) was conducted using an iterative approach. BQSR normally requires true variants to be excluded from error model building. In the absence of prior knowledge of segregating variants, a first round of variant calling was conducted using three different algorithms: GATK UnifiedGenotyper [39], samtools (v. 0.1.18 for both species) [41] and FreeBayes (v. 0.9.8 for crows and v. 0.9.6 for flycatchers) [42]. The single nucleotide polymorphisms (SNPs) detected by the three methods were used as true variants for the BQSR. A first round of BQSR was then run using GATK UnifiedGenotyper exclusively. In the crow, a second round of recalibration was performed but 99.5% of the variants were shared with the first round. On that basis, the calibration was considered to have achieved high consistency and the first round of recalibration was used. For the flycatcher, a second round of BQSR was performed on one population and, because the results were more than 99% identical to the first round, the second round was ignored also in this case. Subsequent variant quality score recalibration was performed with GATK to assign a probability of each SNP being a true variant based on a set of verified variants. Variable sites across populations within species were finally combined and populations

then regenotyped individually using GATK UnifiedGenotyper. As a final conservative filtering step specific to this study, we considered only sites where all individuals within a given population had coverage of at least four.

Nucleotide diversity per 200 kb window was computed on a per-population basis using the Python package pyVCF 0.4.0 and biopython v. 1.68 [43]. Averages were then calculated for each window and species using the mean of both populations; note that per-window diversity levels were strongly correlated between the two collared flycatcher populations (Pearson's $r = 0.99$) and the two hooded crow populations ($r = 0.96$). Windows with fewer than 10 000 sites remaining after coverage based filtering were excluded as was an outlier window in which hooded crow diversity level was far higher (0.0042) than in all other windows (range = 0.0002–0.0023).

Although different software versions were used for GATK, PICARD and FreeBayes in the analyses of the two species, we believe that this has little impact on the results. For example, there were no major changes between v. 2.3.6 and 2.4.9 of GATK. UnifiedGenotyper and FreeBayes only help calibrating GATK.

(c) Data analysis

Collared flycatcher gene annotations were retrieved from ENSEMBL genebuild for release 1.4 of the collared flycatcher genome assembly. Coordinates were then translated to the fAlb15 assembly version. Gene annotations for the hooded crow were obtained from release 100 of GenBank. These annotations were used to estimate coding sequence density. Lineage-specific synonymous substitution rate, d_S, was obtained for the collared flycatcher and was based on data from three-species coding sequence alignments with chicken (*Gallus gallus*) and zebra finch (*Taenopygia guttata*) [44]. After excluding genes with $d_S = 0$ and more than 2 [44], we calculated the average d_S per 200 kb window, weighted by gene length. If the average d_S for a window was above 0.3, d_S was set as missing data. We further obtained data on recombination rate per 200 kb window in the collared flycatcher [22]. These data were originally generated by linkage analysis from the genotyping of a 50 K SNP chip on a large (more than 800 individuals) flycatcher pedigree. Finally, we extracted intergenic GC content as well as the repeat density for each window. We transformed certain candidate explanatory variables to reduce the skewness in their distribution: coding sequence density and d_S were transformed by the square root, and recombination rate was log-transformed to base 10 after adding a constant 1.

We performed a multiple linear regression of collared flycatcher nucleotide diversity (y) against recombination rate (x_1), coding sequence density (x_2), dS (x_3), GC content (x_4) and repeat density (x_5). No interactions were incorporated to avoid over-parametrization:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \in . \qquad (2.1)$$

The underlying assumptions of such linear regression analysis are the lack of heteroscedasticity, multivariate normality and linear relationships between the explanatory variables and the response variable, as well as no collinearity between explanatory variables. In particular, the assumption of no collinearity between explanatory variables appeared problematic. A matrix of pairwise correlation coefficients as well as a correlation tree based on a nested agglomerative method described in [45] is provided as electronic supplementary material.

To handle the problem of collinearity we performed a principal component (PC) regression (PCR), a method derived from principal component analysis (PCA). PCs are calculated using the explanatory variables only. The PCs are then used as predictors for variation in the response variable. It is an efficient way to get around the problem of collinearity between explanatory variables [46].

To specifically investigate the effect of coding sequence density on the extent of linked selection in the flycatcher genome, we regressed nucleotide diversity against recombination rate within gene-rich and gene-poor regions using respectively the highest and the lowest 10% of windows from the distribution of coding sequence density. Because coding sequence density is correlated with GC content, we further regressed nucleotide diversity against recombination rate within GC-rich and GC-poor regions using respectively the highest and the lowest 10% of windows from the distribution of GC content.

3. Results

(a) Levels of genetic diversity

The hooded crow and the collared flycatcher show moderate to moderately high levels of nucleotide diversity with genome-wide averages of $\pi = 0.0039$ (collared flycatcher) and 0.0011 (hooded crow) in the studied populations. Just as observed in many other species, diversity levels vary across the two genomes with π estimates in the 200 kb windows investigated here in the range of 0.0018–0.0060 for collared flycatcher and 0.0002–0.0023 for hooded crows. The chosen window size was considered to be a reasonable trade-off between capturing fine-scale variation in nucleotide diversity and limiting the noise in the estimation of genomic parameters [22]. We retrieved more than 600 Mb (collared flycatcher: 652 Mb; hooded crow: 637 Mb) of conserved synteny between the two species, distributed across all chromosomes. This corresponds to 66% of the flycatcher autosomal assembly (989 Mb) that was used as reference and has scaffolds anchored, ordered and oriented along chromosomes.

As indicated above, a divergence time of at least 25 million years probably means that lineage sorting is completed between the analysed lineages. To test this, we stringently investigated the overlap of segregating sites in the two species by considering the incidence of sites variable in at least one population of flycatchers and one population of crows. Out of 253 303 variable sites in flycatchers, which could be aligned between the two species, only 464 (0.2%) were also variable in crows, confirming that lineage sorting is essentially complete between the two species. This is especially so when considering that any species comparison is bound to include sites polymorphic in both species owing to independently derived mutations, in particular at highly mutable CpG sites.

(b) Correlation of diversity levels between species

Levels of genetic diversity in regions of conserved synteny (200 kb windows; $n = 3259$) of the collared flycatcher and hooded crow genomes were correlated (Spearman's $\rho = 0.407$; $p < 0.0001$; figure 1*a*). Because we analysed more than 60% of the two genomes the investigated regions should provide a representative picture of evolutionary processes affecting genetic diversity in these species. Nevertheless, to exclude biased sampling of genomic regions, we compared the distribution of nucleotide diversity, coding sequence density, recombination rate, d_S, GC content and repeat density between investigated regions and the whole genome (electronic supplementary material, figure S1). The only difference found was a lower repeat density of investigated regions compared with the whole genome, which is probably owing to an expected inverse relationship between repeat density and the ability to identify syntenic regions.

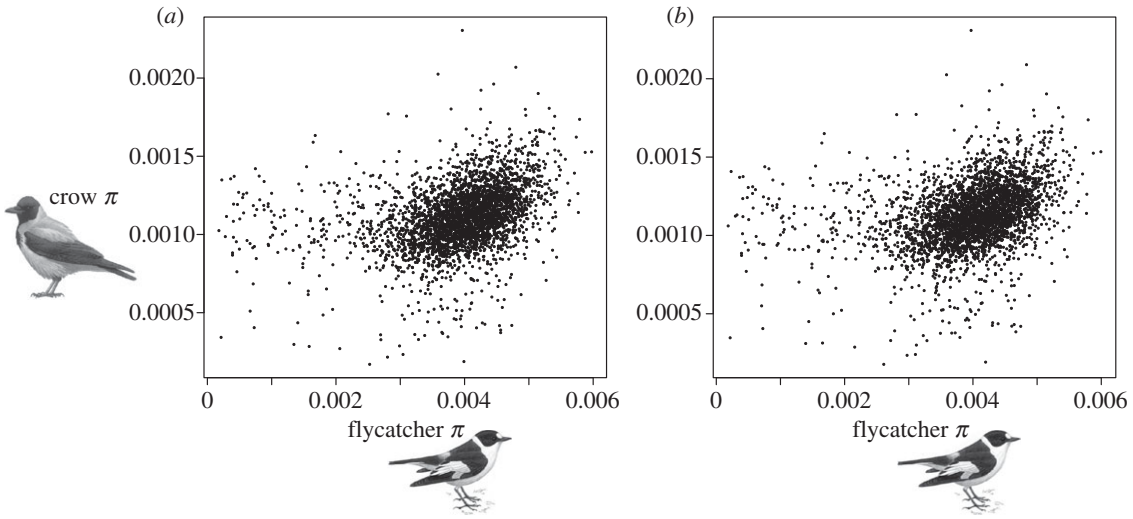

Figure 1. Correlation (Spearman's $\rho = 0.407$) between collared flycatcher and hooded crow nucleotide diversity in 200 kb windows of conserved synteny ($n = 3259$) spread across the genome. (*a*) All sequences, (*b*) excluding coding sequences (Spearman's $\rho = 0.402$).

Table 1. Effect of chromosome length as a covariate of collared flycatcher nucleotide diversity in explaining hooded crow nucleotide diversity. A significant regression equation was found ($F_{2,3256} = 203.3$, *p*-value: $< 2.2 \times 10^{-16}$), with $R^2 = 0.111$.

	t-value	*p*-value
flycatcher diversity	18.93	$<2 \times 10^{-16}$
chromosome length	−1.34	0.162

(c) How can (co)variation in diversity levels be explained?

Previous studies have indicated that genetic diversity within avian genomes varies in relation to chromosome size [47–49]. Together with a stable karyotype, this could potentially lead to an overall correlation between diversity levels in regions of conserved synteny of two species. However, when we regressed hooded crow diversity against collared flycatcher diversity and chromosome length, the effect of chromosome size was not significant (table 1). Another possible factor that could explain a correlation in diversity levels between species is the density of coding sequences (collared flycatcher: mean 0.02 per site, range 0.00–0.15; hooded crow: mean 0.02, range 0.00–0.15) if this density covaries between species and has a large direct effect on 200 kb window-based diversity estimates (given that diversity levels in coding sequences are much lower than in intergenic DNA and introns). When coding sequences were masked, however, the strength of correlation between diversity levels in the two species remained essentially unaltered ($\rho = 0.402$; $p < 0.0001$; figure 1*b*). This therefore suggests that there is some mechanism that affects regional diversity levels in similar ways in syntenic regions of the two genomes.

In order to identify this mechanism, we investigated the driving forces of variation in diversity levels across the avian genome, and given that recombination rate is likely to be a crucial parameter, we focused on the collared flycatcher because pedigree-based recombination rate data are available for this species. We performed a multiple linear regression analysis and, in addition to recombination rate, incorporated coding sequence density as a proxy for the density of targets for selection, d_S as a proxy for the local mutation rate, repeat density and GC content; data for all five explanatory variables were available for 2485 out of the windows used in the flycatcher-crow comparison. All parameters had a significant effect on collared flycatcher genetic diversity and in total explained 28.2% of the variation in genomic diversity (table 2). GC content had the strongest effect (*t*-value = −26.8) followed by recombination rate (*t*-value = 16.0) and d_S (*t*-value = 6.8).

Table 2. Factors explaining collared flycatcher genetic diversity. A significant regression equation was found ($F_{5,2479} = 196.1$, *p*-value: $< 2.2 \times 10^{-16}$), with $R^2 = 0.283$. We analysed a full model including recombination rate coding sequence density, GC content, synonymous substitution rate and repeat density.

	t-value	*p*-value
recombination rate	16.047	$<2 \times 10^{-16}$
coding sequence density	−3.841	1.3×10^{-4}
GC content	−26.772	$<2 \times 10^{-16}$
synonymous substitution rate	6.755	1.8×10^{-11}
repeat density	4.063	5.0×10^{-5}

However, the results of the multiple linear regression need to be interpreted with caution owing to collinearity of the explanatory variables (electronic supplementary material, table S1 and figure S2). In particular, the correlation between GC content and coding sequence density violates the assumption of independence between explanatory variables (Pearson's $r = 0.278$), and thus the respective effects of the two explanatory variables cannot be distinguished and interpreted separately. We therefore performed PCR (figure 2; electronic supplementary material, table S2) to handle the collinearity problem and treat explanatory variables as compounds of their collinearity. This clearly showed that GC content and coding sequence density were tightly linked together, and could therefore not be interpreted separately. PC5, which was mainly governed by GC content, coding sequence density and recombination rate, explained the most of the variance (11.22%). The positive relationship between diversity and recombination rate and the negative relationship between diversity and coding sequence density support a role

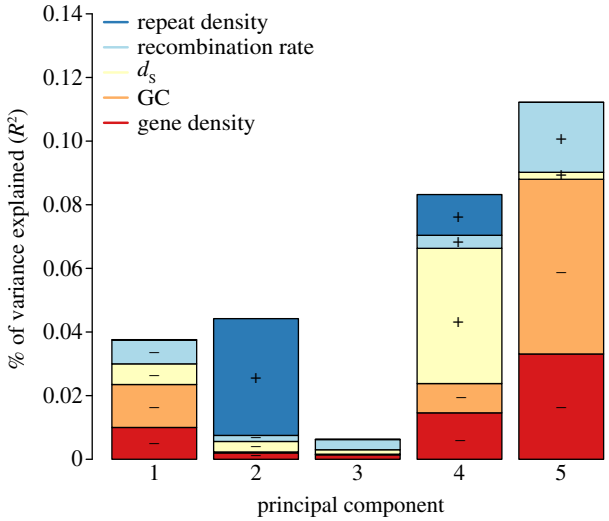

Figure 2. Amount of flycatcher nucleotide diversity explained by different components according to the principal component regression. The analysed explanatory variables are colour-coded according to the key. Plus or minus sign indicates the direction of correlation for individual variables.

of linked selection, because diversity should be most reduced in regions of low recombination and high density of target of selection in a linked selection scenario. The negative relationship between diversity and GC content is in agreement with the observed covariation between coding sequence density and GC content. PC4 explained 8.31% of the variance and was dominated by positive relationships between diversity and d_S, and diversity and repeat density. This would indicate a role of mutation rate variation in explaining variation in diversity levels. PC2 explained 4.42% of the variation and was mostly linked to repeat density. Axes 1 and 3 explained less variance and were difficult to interpret.

To further investigate a role of linked selection on diversity levels we compared genomic regions corresponding to the 10% windows with the lowest coding sequence densities (or GC content) and the 10% windows with the highest coding sequence densities (or GC content). Specifically, we regressed diversity against recombination rate in these two categories of genomic regions to investigate whether the strength of the correlation depends on coding sequence density (or GC content). Recombination rate had a significant effect on diversity in regions with the highest coding sequence density ($F_{1,247} = 22.55$, $p < 0.0001$; $R^2 = 0.08$), but not in regions with the lowest coding sequence density ($F_{1,247} = 0.04$, $p = 0.85$). Similarly, recombination rate had a significant effect on diversity in regions with the highest GC content ($F_{1,247} = 70.44$, $p < 0.0001$; $R^2 = 14.47 = 0.22$), but was reduced in regions with the lowest GC content ($F_{1,247} = 14.47$, $p < 0.0001$; $R^2 = 0.06$). This supports a role of linked selection in governing diversity levels.

4. Discussion

Linked selection affects diversity levels across the genome [1,12]. As predicted by theory, the influence of linked selection has been shown to be affected by several factors, including recombination rate [7,50] and density of targets of selection [6,51]. The extent to which these factors are conserved across species is probably related to general aspects of genome evolution and architecture such as karyotype stability, rate of chromosomal rearrangements and the evolution of base composition. The rate of interchromosomal [16] as well as intrachromosomal rearrangement is low in birds [22,52]. For example, collared flycatcher and zebra finch chromosomes are entirely syntenic and largely collinear [22]. It has been suggested that the stability in genome architecture is associated with stability in genomic features such as recombination rate variation [53]. Indeed, comparisons of broad-scale [22,54] as well as fine-scale (i.e. recombination hot-spots [55]) recombination rates in different avian species indicate that the genomic landscape of recombination rate variation in birds is well conserved. In comparing homologous 1 Mb windows of two distantly related bird species—zebra finch and chicken (*G. gallus*)—Backström *et al.* [54] found that recombination rates were correlated with Spearman's $\rho = 0.50$. Such conservation would promote the build-up over time of correlations between recombination rate and different genomic parameters; a strong correlation observed between recombination rate and base composition represents one such example [56].

We suggest that karyotypic stability and a conserved genomic landscape of recombination rate variation, via the effect they assert on the extent of diversity-reducing linked selection, can at least in part explain the correlation in regional levels of neutral genetic diversity between the collared flycatcher and hooded crow genomes. In the absence of pedigree-based recombination rate data for hooded crow, we cannot formally demonstrate conservation of the recombination landscape compared with collared flycatcher. Crows are difficult to breed in captivity, marked populations cannot easily be followed for many generations in the wild and brood sizes are small, factors that hinder gathering large pedigrees for linkage mapping and associated recombination rate estimation. Moreover, using population-scaled recombination rate data based on the extent of linkage disequilibrium [55] for comparing recombination rate profiles in the two species would be less suitable because linkage disequilibrium is the result of the combined effect of selection and recombination.

In the regression analysis of flycatcher diversity data the PC explaining most of the variance was recombination rate together with density of coding sequence, consistent with linked selection. The second strongest PC included mainly d_S and repeat density. With d_S considered a proxy for the neutral mutation rate and with some evidence for a link between open chromatin, mutation rate and the abundance of transposable elements [11], this would indicate that mutation rate variation contributes to regional variation in genetic diversity. Although theoretically expected (given $\theta = 4 N_e \mu$), there is mixed evidence from empirical studies of a relationship between diversity and mutation rate, possibly because covariation of several genomic variables blurs potential effects of mutation rate variation on diversity. Nevertheless, a stable avian karyotype could allow for a stable genomic environment, leading not only to a stable recombination landscape, but also to a conserved landscape of mutation rate and chromatin structure. Further work should be devoted to analyses of the relationship between mutation and diversity. In the long term, direct estimates (in contrast to indirect estimates obtained from diversity or divergence data) of local mutation rates from pedigrees or mutation accumulation lines are likely to become available and will be quite informative in this respect.

In summary, together with similar results obtained in a comparison of *Drosophila melanogaster* and *D. simulans* [57], our study is one of the first to demonstrate a genome-wide

correlation in regional levels of genetic diversity in two lineages long after sorting of ancestral variation. This covariation is seen despite that very different selection pressures (e.g. on life history, ecology, morphology and behaviour) are likely to have operated in the two investigated avian lineages for millions of years. We suggest that the correlation can be explained by a similar genomic architecture of factors governing diversity levels through linked selection, namely karyotypic stability and a conserved recombination rate landscape. More generally, karyotype stability may imply a conserved genomic environment, such that conservation in other factors such as mutation rate variation reinforces the correlation. Our observations imply that genetic diversity is to some extent predictable.

Data accessibility. Accession numbers to all sequence data analysed are provided in the original reports of crow and flycatcher polymorphism data. Genomic and population genetic data used in the analyses are provided in electronic supplementary material, table S3.

Authors' contributions. L.D. made all data analyses (with input from C.F.M. and J.W.). L.D. and H.E. wrote the paper. N.V. provided crow polymorphism data. C.M.B. and R.B. made initial processing of crow and flycatcher sequence data, respectively. H.E. and J.W. conceived of and designed the project. H.E. supervised the project.

Competing interests. There is no competing interest to be declared.

Funding. Funding was obtained from the European Research Council, the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

Acknowledgements. We thank Juha Merilä and two anonymous reviewers for helpful suggestions for improvement.

References

1. Ellegren H, Galtier N. 2016 Determinants of genetic diversity. *Nat. Rev. Genet.* **17**, 422–433. (doi:10.1038/nrg.2016.58)
2. Haasl RJ, Payseur BA. 2016 Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Mol. Ecol.* **25**, 5–23. (doi:10.1111/mec.13339)
3. Cruickshank TE, Hahn MW. 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157. (doi:10.1111/mec.12796)
4. Frankham R. 2015 Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Mol. Ecol.* **24**, 2610–2618. (doi:10.1111/mec.13139)
5. Charlesworth B. 2009 Effective population size and patterns of molecular evolution and variation. *Nat. Rev. Genet.* **10**, 195–205. (doi:10.1038/nrg2526)
6. Gossmann TI, Woolfit M, Eyre-Walker A. 2011 Quantifying the variation in the effective population size within a genome. *Genetics* **189**, 1389–1402. (doi:10.1534/genetics.111.132654)
7. Begun DJ, Aquadro CF. 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* **356**, 519–520. (doi:10.1038/356519a0)
8. Consortium TGP. 2015 A global reference for human genetic variation. *Nature* **526**, 68–74. (doi:10.1038/nature15393)
9. Fijarczyk A, Babik W. 2015 Detecting balancing selection in genomes: limits and prospects. *Mol. Ecol.* **24**, 3529–3545. (doi:10.1111/mec.13226)
10. Hodgkinson A, Eyre-Walker A. 2011 Variation in the mutation rate across mammalian genomes. *Nat. Rev. Genet.* **12**, 756–766. (doi:10.1038/nrg3098)
11. Makova KD, Hardison RC. 2015 The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.* **16**, 213–223. (doi:10.1038/nrg3890)
12. Cutter AD, Payseur BA. 2013 Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat. Rev. Genet.* **14**, 262–274. (doi:10.1038/nrg3425)
13. Charlesworth B, Morgan MT, Charlesworth D. 1993 The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**, 1289–1303.
14. Gillespie JH. 2000 Genetic drift in an infinite population: the pseudohitchhiking model. *Genetics* **155**, 909–919.
15. Smith JM, Haigh J. 1974 The hitch-hiking effect of a favourable gene. *Genet. Res.* **23**, 23–35. (doi:10.1017/S0016672300014634)
16. Ellegren H. 2010 Evolutionary stasis: the stable chromosomes of birds. *Trends Ecol. Evol.* **25**, 283–291. (doi:10.1016/j.tree.2009.12.004)
17. Gregory T. 2016 Animal genome size database. See http://www.genomesize.com/.
18. Shields G. 1982 Comparative avian cytogenetics: a review. *Condor* **84**, 45–58. (doi:10.2307/1367820)
19. Roslik G, Kryukov A. 2001 Karyological study of some corvine birds (Corvidae, Aves). *Russ. J. Genet.* **37**, 796–806. (doi:10.1023/A:1016703127516)
20. Bian X, Li Q, Ning S. 1991 Studies on the karyotypes of birds III. 10 species of warblers and 4 species of flycatchers (Aves). *Zool. Res.* **12**, 215–220.
21. Ellegren H *et al.* 2012 The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760. (doi:10.1038/nature11584)
22. Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, Mugal CF, Olason P, Ellegren H. 2014 A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. *Mol. Ecol.* **23**, 4035–4058. (doi:10.1111/mec.12810)
23. Poelstra JW *et al.* 2014 The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**, 1410–1414. (doi:10.1126/science.1253226)
24. Poelstra JW, Vijay N, Hoeppner MP, Wolf JBW. 2015 Transcriptomics of colour patterning and coloration shifts in crows. *Mol. Ecol.* **24**, 4617–4628. (doi:10.1111/mec.13353)
25. Jarvis ED *et al.* 2014 Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331. (doi:10.1126/science.1253451)
26. Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015 A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569–573. (doi:10.1038/nature15697)
27. Jønsson KA, Fabre P-H, Kennedy JD, Holt BG, Borregaard MK, Rahbek C, Fjeldså J. 2016 A supermatrix phylogeny of corvoid passerine birds (Aves: Corvides). *Mol. Phylogenet. Evol.* **94**A, 87–94. (doi:10.1016/j.ympev.2015.08.020)
28. Vijay N, Bossu CM, Poelstra JW, Weissensteiner MH, Suh A, Kryukov AP, Wolf JBW. 2016 Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. *Nat. Commun.* **7**, 13195. (doi:10.1038/ncomms13195)
29. Brommer JE, Gustafsson L, Pietiäinen H, Merilä J. 2004 Single-generation estimates of individual fitness as proxies for long-term genetic contribution. *Am. Nat.* **163**, 505–517. (doi:10.1086/382547)
30. Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. 2013 Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genet.* **9**, e1003942. (doi:10.1371/journal.pgen.1003942)
31. Nadachowska-Brzyska K, Burri R, Smeds L, Ellegren H. 2016 PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol. Ecol.* **25**, 1058–1072. (doi:10.1111/mec.13540)
32. Wolf JBW, Bayer T, Haubold B, Schilhabel M, Rosenstiel P, Tautz D. 2010 Nucleotide divergence vs. gene expression differentiation: comparative transcriptome sequencing in natural isolates from the carrion crow and its hybrid zone with the hooded crow. *Mol. Ecol.* **19**, 162–175. (doi:10.1111/j.1365-294X.2009.04471.x)
33. Hudson RR, Coyne JA. 2002 Mathematical consequences of the genealogical species concept. *Evolution* **56**, 1557–1565. (doi:10.1111/j.0014-3820.2002.tb01467.x)
34. Harris RS. 2007 Improved pairwise alignment of genomic DNA. PhD thesis, Pennsylvania State University, State College, PA.

35. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002 The human genome browser at UCSC. *Genome Res.* **12**, 996–1006. (doi:10.1101/gr.229102. Article published online before print in May 2002)
36. Tang H, Krishnakuar V, Li J. 2015 jcvi: JCVI utility libraries. Zenodo. (doi:10.5281/zenodo.31631)
37. Burri R *et al.* 2015 Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* **25**, 1656–1665. (doi:10.1101/gr.196485.115)
38. Li H, Durbin R. 2009 Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760. (doi:10.1093/bioinformatics/btp324)
39. DePristo MA *et al.* 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498. (doi:10.1038/ng.806)
40. McKenna A *et al.* 2010 The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303. (doi:10.1101/gr.107524.110)
41. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
42. Garrison E, Marth G. 2012 Haplotype-based variants detection from short-read sequencing. *ArXiv* 1207.3907.
43. Cock PJ *et al.* 2009 Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423. (doi:10.1093/bioinformatics/btp163)
44. Bolívar P, Mugal CF, Nater A, Ellegren H. 2016 Recombination rate variation modulates gene sequence evolution mainly via GC-biased gene conversion, not Hill–Robertson interference, in an avian system. *Mol. Biol. Evol.* **33**, 216–227. (doi:10.1093/molbev/msv214)
45. Kaufman L, Rousseeuw PJ. 1990 *Finding groups in data: an introduction to cluster analysis*. New York, NY: Wiley.
46. Mugal CF, Nabholz B, Ellegren H. 2013 Genome-wide analysis in chicken reveals that local levels of genetic diversity are mainly governed by the rate of recombination. *BMC Genomics* **14**, 86. (doi:10.1186/1471-2164-14-86)
47. Manthey JD, Klicka J, Spellman GM. 2015 Chromosomal patterns of diversity and differentiation in creepers: a next-gen phylogeographic investigation of *Certhia americana*. *Heredity* **115**, 165–172. (doi:10.1038/hdy.2015.27)
48. Callicrate T, Dikow R, Thomas JW, Mullikin JC, Jarvis ED, Fleischer RC. 2014 Genomic resources for the endangered Hawaiian honeycreepers. *BMC Genomics* **15**, 1098. (doi:10.1186/1471-2164-15-1098)
49. ICGSC. 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716. (doi:10.1038/nature03154)
50. Corbett-Detig RB, Hartl DL, Sackton TB. 2015 Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol.* **13**, e1002112. (doi:10.1371/journal.pbio.1002112)
51. Slotte T. 2014 The impact of linked selection on plant genomic variation. *Brief. Funct. Genomics* **13**, 268–275. (doi:10.1093/bfgp/elu009)
52. Volker M, Backstrom N, Skinner BM, Langley EJ, Bunzey SK, Ellegren H, Griffin DK. 2010 Copy number variation, chromosome rearrangement, and their association with recombination during avian evolution. *Genome Res.* **20**, 503–511. (doi:10.1101/gr.103663.109)
53. Mugal CF, Arndt PF, Ellegren H. 2013 Twisted signatures of GC-biased gene conversion embedded in an evolutionary stable karyotype. *Mol. Biol. Evol.* **30**, 1700–1712. (doi:10.1093/molbev/mst067)
54. Backström N *et al.* 2010 The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res.* **20**, 485–495. (doi:10.1101/gr.101410.109)
55. Singhal S *et al.* 2015 Stable recombination hotspots in birds. *Science* **350**, 928–932. (doi:10.1126/science.aad0843)
56. Nabholz B, Kunstner A, Wang R, Jarvis ED, Ellegren H. 2011 Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol. Biol. Evol.* **28**, 2197–2210. (doi:10.1093/molbev/msr047)
57. Langley CH *et al.* 2012 Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics* **192**, 533. (doi:10.1534/genetics.112.142018)