DR. NAGARJUN VIJAY (Orcid ID : 0000-0003-3955-3369)
DR. RETO BURRI (Orcid ID : 0000-0002-1813-0079)
PROF. HANS ELLEGREN (Orcid ID : 0000-0002-5035-1736)

Article type : Original Article

Genome-wide patterns of variation in genetic diversity are shared among populations, species and higher order taxa

Nagarjun Vijay[1,2], Matthias Weissensteiner[1,5], Reto Burri[1,3], Takeshi Kawakami[1,4], Hans Ellegren[1], Jochen B. W. Wolf[1,5]

[1] Department of Evolutionary Biology and SciLifeLab, Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden

[2] Lab of Molecular and Genomic Evolution, Department of Ecology and Evolutionary Biology, College of Literature, Science, and the Arts, University of Michigan, Ann Arbor, MI, USA

[3] Department of Population Ecology, Friedrich Schiller University Jena, Dornburger Strasse 159, 07743 Jena, Germany

[4] Department of Animal and Plant Sciences, University of Sheffield, S10 2TN, United Kingdom

[5] Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilians-Universität München, Grosshaderner Strasse 2, 82152 Planegg-Martinsried , Germany

Corresponding authors: nagarjun.vijay@ebc.uu.se; j.wolf@bio.lmu.de

Keywords: speciation genetics, linked selection, recombination rate, genetic diversity, genetic

ABSTRACT

Genome-wide screens of genetic variation within and between populations can reveal signatures of selection implicated in adaptation and speciation. Genomic regions with low genetic diversity and elevated differentiation reflective of locally reduced effective population sizes (N_e) are candidates for barrier loci contributing to population divergence. Yet, such candidate genomic regions need not arise as a result of selection promoting adaptation or advancing reproductive isolation. Linked selection unrelated to lineage-specific adaptation or population divergence can generate comparable signatures. It is challenging to distinguish between these processes, particularly when diverging populations share ancestral genetic variation. In this study, we took a comparative approach using population assemblages from distant clades assessing genomic parallelism of variation in N_e. Utilizing population-level polymorphism data from 444 re-sequenced genomes of three avian clades spanning 50 million years of evolution we tested whether population genetic summary statistics reflecting genome-wide variation in N_e would co-vary among populations within clades, and importantly, also among clades where lineage sorting has been completed. All statistics including population-scaled recombination rate (ρ), nucleotide diversity (π) and measures of genetic differentiation between populations (F_{ST}, PBS, d_{xy}) were significantly correlated across all phylogenetic distances. Moreover, genomic regions with elevated levels of genetic differentiation were associated with inferred peri-centromeric and sub-telomeric regions. The phylogenetic stability of diversity landscapes and stable association with genomic features support a role of linked selection not necessarily associated with adaptation and speciation in shaping patterns of genome-wide heterogeneity in genetic diversity.

INTRODUCTION

Understanding the processes governing heterogeneity of genome-wide diversity has been a long-standing goal in evolutionary genetics (Ellegren and Galtier 2016), and is of central importance to adaptation and speciation research (Seehausen et al. 2014; Wolf and Ellegren 2017). A plethora of recent studies characterising genetic variation of diverging natural populations in a taxonomically diverse set of species identified strong heterogeneity in the genome-wide distribution of genetic diversity, both within and between populations (e.g. in sunflowers (Renaut et al. 2013), monkey flowers (Puzey et al. 2017), stickleback fish (Roesti et al. 2015), rabbits (Carneiro et al. 2014) or

birds (Ellegren et al. 2012, Poelstra et al. 2014)). Despite commonality in patterns seen across this wide range of taxa, elucidating the underlying processes remains challenging (Wolf and Ellegren 2017).

Regions of reduced genetic diversity generally coinciding with elevated levels of genetic differentiation (Charlesworth 1998) can be interpreted in the context of adaptation and speciation under conditions of gene flow (Nosil and Feder 2013). Building on the idea of a 'genic view of speciation' (Wu 2001), barrier loci experiencing divergent selection contribute to a reduction of gene flow between populations (i.e., reduced effective migration rate (m_e) relative to gross migration rate (m) (Abbott et al. 2013)). However, recombination decouples the locus under divergent selection from neighboring genetic variation. As a consequence, effective migration rates will not only vary across the genome as a function of the strength of selection (s), but also due to recombination rate (r). Effective migration will be most strongly be reduced by selection at the causative locus and increases as a function of genetic distance to levels experienced by neutral genetic variation (at equilibrium $m_e=m/(1+s/r)$, (Barton and Bengtsson 1986)). Assuming neutrality, empirical information on genome-wide migration rate under mutation-drift equilibrium can be obtained from measures of genetic differentiation, usually $F_{ST}\sim1/(1+N_e(m+\mu))$. Genome scans assaying local levels of genetic differentiation along the genome may additionally allow identifying regions under selection (Lewontin and Krakauer 1973). Positive selection will reduce local levels of genetic diversity, and hence N_e, resulting in increased levels of F_{ST} (see also (Cruickshank and Hahn 2014)). Divergent selection opposing gene flow between populations will further increase regional genetic differentiation by preventing homogenizing admixture (reducing m_e). Regions of the genome with elevated levels of genetic differentiation and reduced levels of genetic diversity are thus often regarded as candidates for hosting barrier loci subject to divergent selection and refractory to the homogenizing process of gene flow ('speciation islands') (Nosil and Feder 2013). Although often framed in the context of ecological speciation (Nosil and Feder 2013), barrier loci refer to any genetic element conveying ecological, sexual, pre- or postzygotic reproductive isolation (Wolf et al. 2010). The cumulative effect of multiple barrier loci is eventually expected to transition to genome-wide barriers, ultimately promoting speciation (Barton 1983; Abbott et al. 2013).

However, divergent selection promoting lineage-specific adaptation or reproductive isolation under conditions of gene flow is not the only process introducing heterogeneity in N_e across the genome.

Any form of selection that reduces genetic diversity will result in comparable signatures of genome-wide heterogeneity in N_e. Selection reducing diversity not only at sites under selection, but also at linked neutrally evolving sites, is collectively referred to as linked selection. This includes both positive selection (Smith and Haigh 1974) and negative (background) selection (Charlesworth et al., 1993; Charlesworth 1994). Although these two selective mechanisms are fundamentally different, it is difficult to discern their effect on genetic diversity and differentiation (Stephan 2010). Linked selection is expected to be most pronounced in regions of low recombination and high target (gene) density, and has been shown to significantly affect heterogeneity in levels of genetic diversity across a broad range of organisms (Nachman and Payseur 2012; Cutter and Payseur 2013; Slotte 2014; Burri et al. 2015). Genomic regions subject to linked selection are not only depleted of genetic diversity ($\theta \sim N_e\mu$), but also experience accelerated lineage sorting resulting in increased levels of relative genetic differentiation (F_{ST}) (Renaut et al. 2013; Cruickshank and Hahn 2014). Relating patterns of genetic variation and differentiation to the underlying process is further complicated by additional intrinsic and extrinsic factors such as mutation rate variation or demographic perturbation (Strasburg et al. 2012).

Several ways forward have been suggested to differentiate between linked selection universally acting in all populations from lineage-specific selection promoting adaptation and speciation. *Functional validation* of candidate barrier loci flagged during genome scans provides valuable, independent information on the plausibility of divergent selection opposing gene flow in a given population-specific context (Kronforst and Papa 2015). *Theoretical models* provide useful null expectations to compare with empirical patterns (Bank et al. 2014). *Experimental evolution* studies (Dettman et al. 2007) or manipulative experiments in natural populations (Soria-Carrasco et al. 2014) allow the link between the nature of selection and genomic patterns of genetic diversity to be studied under controlled conditions. *Micro-level comparative population approaches* leveraging information from spatiotemporal contrasts between populations ('speciation continuum' (Mallet et al. 2007; Powell et al. 2013; Seehausen et al. 2014)) help disentangle the effects of linked selection unrelated to speciation (e.g. background selection) from those thought to contribute to reproductive isolation in the face of gene flow (e.g. divergent selection) (Wolf and Ellegren 2017). This includes the use of natural hybrids (Barton 1983; Gompert and Buerkle 2011) or crosses generated in the lab (Seehausen et al. 2014). Within species and among closely related species, however, a substantial fraction of genetic variation is shared by ancestry, impeding inference.

Here we propose a *macro-level comparative approach* extending comparisons of genome-wide diversity beyond closely related taxa to phylogenetically distant clades, where lineage sorting has long been completed. This controls for the effect of shared recent ancestry, recent or ongoing gene flow between clades. Genomic parallelism in patterns of genetic diversity across such large evolutionary distances cannot be explained by processes involving selection on a set of specific genes for each lineage. Instead, it is expected that genomic parallelism is mediated by universal processes shared in syntenic regions with similar genomic properties among clades.

One candidate parameter to affect genetic diversity ($\theta=4N_e\mu$) of syntenic regions similarly among clades is the mutation rate μ, which is known to vary across the genome (Hodgkinson and Eyre-Walker 2011). However, support for a role of mutation rate in modulating the level of genetic variation and differentiation across the genome is limited (Cutter and Payseur 2013). While some studies found a contribution (Dutoit et al. 2017; Smith and Eyre-Walker 2017), genetic diversity is generally only weakly associated with proxies for mutation rate (Cutter and Payseur 2013; Vijay et al. 2016). Another parameter that can affect genetic diversity is recombination rate which is reportedly conserved at broad-scale between clades (Roesti et al. 2012; Auton et al. 2012; Kawakami et al. 2014; Tine et al. 2014; Burri et al. 2015; Singhal et al. 2015). With little evidence for recombination-associated mutation (and hence r~μ) (Cutter and Payseur 2013) any form of linked selection, where the local reduction in N_e through selection is contingent on the rate of local recombination, is thus a prime candidate for explaining shared heterogeneity in genetic variation among clades (Cutter and Payseur 2013).

A *macro-level comparative* perspective on the genome-wide variation of genetic diversity is implicit, though not the main focus, of recent work by Van Doren et al. (2017) and Dutoit et al. (2017) comparing summary statistics of genetic diversity between stonechats and flycatchers, and between flycatchers and crows, respectively. Here, we assess the contribution of linked selection in shaping genome-wide landscapes of genetic diversity and differentiation across a wide range of evolutionary time-scale ranging from few thousand to approximately 50 million years of evolution. Given the global conservation of recombination landscape for tens of millions of years among avian lineages (Singhal et al. 2015), it is expected that linked selection mediated by recombination constitutes an important component for the concerted evolution of heterogeneity in genome-wide

diversity. Note that linked selection resulting in genomic parallelism between clades includes background selection as well as positive selection acting repeatedly on orthologous loci among clades. We, therefore, predict that summary statistics reflective of N_e not only co-vary among populations of closely related taxa, but also are correlated among clades. Moreover, assuming karyotypic stability, we would expect genomic regions with locally reduced N_e by linked selection to be stably associated with chromosomal features of suppressed recombination such as peri-centromeric or sub-telomeric regions.

To empirically address this expectation, we used publicly available genome re-sequencing data from several populations or (sub)-species of three distantly related clades of avian species complexes - Darwin's finches, *Ficedula* flycatchers and *Corvus* crows (**Table S1**) - with split times beyond the expected time for complete lineage sorting (**Fig. 1, Supporting Information**). For each population and species comparison within clades, we quantified a set of genetic summary statistics in syntenic windows of 50 kb in size. Summary statistics were chosen to be reflective of the local effective population size (N_e) of a genomic region: population-scaled recombination rate ρ ($\sim N_e r$), nucleotide diversity π ($\sim N_e \mu$), genetic differentiation expressed as F_{ST} ($\sim 1/(1+N_e (m+\mu))$ (where mutation rate μ can generally be neglected if migration rate $m >> \mu$), the related population branch statistic (PBS) accounting for non-independence of population comparisons, and d_{xy} ($\sim N_e \mu + \mu t$) reflecting the average number of nucleotide substitutions between populations. The only parameter shared by these statistics is N_e; hence, co-variation of all statistics in syntenic regions would indicate selection affecting local N_e alike in the investigated populations.

MATERIAL AND METHODS

Clades

We chose populations and (sub)-species from three phylogenetically divergent clades: Darwin's finches of the genera *Geospiza, Certhidea and Platyspiza.,* flycatchers of the genus *Ficedula* (*F. albicollis, F. hypoleuca, F. semitorquata* and *F. speculigera)* and crows of the genus *Corvus* including the American crow *C. brachyrhynchos* and several taxa from the *Corvus (corone) spp.* species complex (Vijay et al. 2016). Functionally annotated genome assemblies with high sequence contiguity are available for one representative each of *Ficedula* flycatchers (*F. albicollis,* genome size: 1.13, scaffold / contig N50= 6.5 Mb / 410 kb, National Center for Biotechnology Information (NCBI) accession number: GCA_000247815.2; (Ellegren et al. 2012); new chromosome build

(Kawakami et al. 2014)) and for one hooded crow specimen (*Corvus (corone) cornix*, genome size: 1.04Gb, scaffold / contig N50 = 16.4 Mb / 94 kb, NCBI accession number: GCA_000738735.1; (Poelstra et al. 2014, 2015)). The assembly of the medium ground finch *G. fortis* is of comparable size (1.07 Gb) and the least contiguous among the three both at the scaffold and contig level (scaffold / contig N50 = 5.3 Mb / 30 kb, NCBI accession number: GCA_000277835.1; (Rands et al. 2013)).

In all three clades, it has been suggested that shared genetic variation between (sub)-species within clades resulted from incomplete lineage sorting of ancestral polymorphisms, regardless of whether populations were connected by recent gene flow or not (Lamichhaney et al. 2015; Burri et al. 2015; Vijay et al. 2016). However, shared polymorphism is highly unlikely among clades because of their phylogenetic distance. Phylogenetic relationships and divergence time estimates between representatives of all three clades and zebra finch (*Taenopygia guttata*) as shown in **Fig. 1** has been extracted as the consensus of 10,000 phylogenetic reconstructions from Jetz et al. (2012; 2014) using the tree of 6670 taxa with sequence information by Ericson et al. (2006) as the backbone (http://birdtree.org/). This places the separation between Corvoidea (crows) and Passerida (Darwin's finches and flycatchers) at over 50 million years. Assuming a range in generation time between six years for hooded crows (Vijay et al. 2016), five years for Darwin's finches (Grant and Grant 1992) and two years for flycatchers (Brommer et al. 2004) this corresponds to at least 8-25 million generations. With an estimated long-term N_e of 200,000 for flycatchers and crows (Wolf et al. 2010; Nadachowska-Brzyska et al. 2013; Vijay et al. 2016) and considerably less for Darwin's finches (N_e = 6,000 to 60,000, (Lamichhaney et al. 2015)) this yields a minimum range of 40-125 N_e generations as time to the most common ancestor. This is clearly beyond the expected time for complete lineage sorting (9-12 N_e generations; (Hudson and Coyne 2002)), and clades are thus not expected to share ancestral polymorphism. The same consideration holds for the split between flycatcher and Darwin's finches assuming approximately 45 million years of divergence (**Fig. 1**). Even assuming an earlier, minimal age estimate of the split between Corvoidea and Passerida in the order of 25 million years ago (Jarvis et al. 2014; Prum et al. 2015; Jønsson et al. 2016) and a split between flycatchers and finches at 19 million years (Singhal et al. 2015) gives split times beyond 12 N_e generations suggesting complete lineage sorting for neutral genetic variation.

Establishing homology among genomes

Homologous regions between genomes were identified in order to quantify the degree to which genetic diversity, recombination and genetic differentiation landscapes are conserved between species. To ensure comparability across all three clades in the most efficient way, we chose to lift-over coordinates of 50 kb non-overlapping windows from the genomes to the independent, well maintained high quality zebra finch reference genome (Hubbard et al. 2002). Lift-over is the process of transferring the positions along one genome to another genome based on whole genome alignments. This approach assumes a high degree of synteny among species, which is justified given the evolutionary stasis of chromosomal organisation in birds across more than 100 million years of evolution (Ellegren 2010). Performing a base by base lift-over can lead to partial loss of regions within a window as well as merging of non-adjacent windows. While sequencing reads of one species can be mapped to the genome of another species to identify variants, this strategy cannot be confidently extended beyond 5-15% sequence divergence without introducing read mapping bias (Vijay et al. 2013; Shafer et al. 2016). To avoid such errors we estimated the statistics for each species in windows prior to the lift-over. Converting the coordinates of genomes from multiple different species into one single coordinate system allows for straightforward comparison of all statistics derived from the original polymorphism data (in Variant Call Format or vcf).

Whole genome alignments between species can be represented in the form of chain files that record the links between orthologous regions of the genome. We downloaded chain files from the UCSC website (https://genome.ucsc.edu/) to transfer the coordinates in bed format from flycatcher and Darwin's Finch genomes onto the zebra finch genome using the program liftOver (Kuhn et al. 2007). For the crow genome where no chain files were available, we first aligned the crow genome to the flycatcher genome using LASTZ (Harris 2007) to obtain a .psl file which was subsequently converted to a chain file using JCVI utility libraries (Tang et al. 2015). This chain file was then used to transfer the crow coordinates to zebra finch coordinates (via flycatcher) using the liftOver utility (Hinrichs et al. 2006).

Orthology could be established for a large proportion of the original genomes. Depending on parameter settings controlling stringency ('minmatch') and cohesion ('minblocks') percent recovery ranged from as little as 13% to over 90% (**Fig. S1, Table S2**). To find an optimal combination of parameter values and to validate lift-over quality, we made use of the fact that GC content in orthologous regions of avian genomes is expected to be strongly conserved across long evolutionary

distances (Weber et al. 2014). We calculated GC content in 50 kb windows from the three different assemblies and compared these values to the GC content at the new, orthologous positions lifted over to the zebra finch genome. Pearson's correlations were high across a broad set of parameter values in all clades ranging from 0.83-0.97. While liftOver is able to transfer the coordinates from the focal genome onto positions along the zebra finch genome these new positions do not retain the window structure from the original genomes. To be able to compare population genetic summary statistics between species in orthologous windows, we defined 50 kb windows along the zebra finch genome. For each window, we then calculated a mean value across all regions that were lifted over and overlapped a given window. To ensure that this procedure of calculating means did not unduly influence comparability across species, we compared the values of GC content from each of the focal genomes after taking the mean across overlapping regions to the GC content in the zebra finch genomic windows. Although correlation coefficients were lower than those seen directly after liftOver, they still exceeded 0.78, 0.82, 0.82 for Darwin's finch, flycatcher and crow respectively across a broad 'minmatch' and 'minblock' parameter space (**Fig. S1, Table S2**). The high correlation of GC content across the liftOver steps suggests that the lift over procedure of moving the windows from one genome assembly to another was reliable at the window size being evaluated. Finally, an optimal combination of stringency, cohesion and percent recovery was chosen on the basis of the (visually inferred) inflection point of the relationship between GC correlation and recovery (**Fig. S1)**.

It could be seen that certain regions of the genome were systematically more susceptible to drop out during liftOver than others for all clades (**Fig. S2**). In particular, regions located on scaffolds that have not been linked to any specific chromosome and those that have not been placed at a particular position along a chromosome were more difficult to lift-over than other regions of the genome. Hence, for the purpose of this study, we have excluded these regions in all subsequent analyses. To ensure that liftOver did not introduce a bias in the regions being analysed, we compared the GC content distribution of the regions that could be lifted over at different values of the "minmatch" parameter (**Fig. S3)**. No clear evidence of bias with regard to GC content of the successfully lifted over regions emerged.

Datasets

We compiled the following publicly available population re-sequencing datasets for the three clades

(**Table S1**). Populations with less than three individuals were excluded in all species.

1. Crows in the genus *Corvus* (124 genomes resequenced, 55 population comparisons within and between 2 focal species, the American crow *C. brachyrhynchos* and various (sub)-species and populations within the *C. (corone) spp.* complex). Population genetic summary statistics including genetic diversity (π), population recombination rate (ρ), genetic differentiation (F_{ST}, PBS, d_{xy}) across the European crow hybrid zone have been characterised using high coverage whole genome re-sequencing data of 60 individuals samples in a 2×2 population design between carrion crows *(Corvus (corone) corone)* and hooded crows *(C. (c.) cornix)* (Poelstra et al. 2014). This study has been followed by a broader sampling regime with a total of 118 crows from the *Corvus (c.) spp.* species complex including a parallel hybrid zone in Russia between *C. (c.) cornix* and *C. (c.) orientalis,* a contact zone between the latter and *C. (c.) pectoralis* and numerous other allopatric populations (Vijay et al. 2016). The system is relatively young, such that 12% of segregating genetic variation has been estimated to be shared between Eurasian and American crows (*C. brachyrhynchos)* (Vijay et al. 2016) which split at approximately 3 million years ago (Jønsson et al. 2016). F_{ST} and d_{xy} ranged from 0.016-0.486 and 0.0015-0.0018 respectively. A broad range in π (0.0010-0.0033) and Tajima's D (0.5895 to -1.974) suggests perturbation by population-specific demographic histories.

2. *Ficedula flycatchers* (200 genomes resequenced with 30 population comparisons across the 4 focal species *F. albicollis, F. hypoleuca, F. semitorquata and F. speculigera* and two outgroup species *F. parva* and *F. hyperythra).* Species diverged approximately 2 million years ago and populations differ slightly in genome-wide levels of differentiation (π: 0.0029-0.0039). A total of 30 population comparisons within and across species provide a broad contrast across a spectrum of genome-wide differentiation (F_{ST}: 0.012-0.981 and d_{xy}: 0.0031-0.0050) (see (Burri et al. 2015)).

3. *Darwin's finches* (120 genomes resequenced, 44 population comparisons across the 6 focal species *Geospiza conirostris, Geospiza difficilis, Camarhynchus pallidus, Certhidea fusca, Certhidea olivacea* and *Pinaroloxias inornata).* The differentiation landscape of Darwin's finches has been studied using whole genome re-sequencing data and has been instrumental in the identification of adaptive loci associated with beak shape evolution (Lamichhaney et

al. 2015). This set of populations across several species differs slightly in genome-wide levels of diversity (π: 0.0003-0.0012, see (Lamichhaney et al. 2015)). Species are estimated to share common ancestry ~1.5 million years ago, yielding 44 population comparisons ranging across a broad spectrum of genome-wide differentiation (F_{ST}: 0.192-0.897) and divergence (d_{xy}: 0.0022-0.0047).

Genetic diversity data

In all three study systems segregating genetic variation and related summary statistics have been characterized in non-overlapping windows across the genome using similar strategies based on the Genome Analysis Toolkit GATK (DePristo et al. 2011) (see **Table S3** for methodological comparison and consult individual studies for additional details). We used the final set of variant calls from each individual to calculate a set of summary statistics. vcf (Variant Call Format) files were obtained from Lamichaney et al. (2015) for Darwin's finches, Burri et al. (2015) for flycatchers and Vijay et al. (2016) for crows. Each of the statistics were calculated in 50 kb windows for all scaffolds longer than 50 kb.

Population recombination rate (ρ) and nucleotide diversity (π)

To generate an estimate of the population-scaled recombination rate in Darwin's finches ρ we followed the approach described in Vijay et al. (2016). In brief, we used LDhelmet (Chan et al., 2012) on genotype data phased with fastPHASE (Scheet and Stephens 2006). The required mutation matrix was approximated from zebra finch substitution rates following Singhal et al. (2015). Population recombination rate data for crows and flycatchers were estimated using the same approach and were extracted from Vijay et al. (2016) and Kawakami et al. (2017), respectively. Pairwise nucleotide diversity π was calculated from the .vcf files using the R package Hierfstat. The number of usable invariant sites were identified based on per base pair sequencing coverage of individuals to use only those sites that are covered by at least 5 reads in more than half of the individuals in each population.

Genetic differentiation (F_{ST}, PBS, d_{xy})

F_{ST} was estimated using Weir and Cockerham's estimator based on genotypes from the .vcf files using the procedure implemented in the Hierfstat package (Goudet 2005) as the ratio of the average of variance components. To avoid pseudo-replicated populations comparisons we also calculated

lineage-specific F_{ST} in the form of population branch statistics (PBS) using the formula PBS_Pop1=(-log(1-F_{ST}(Pop1_Pop2)+(-log(1-F_{ST}(Pop1_Pop3)))-log(1-F_{ST}(Pop2_Pop3)/2. d_{xy} following the definition by Nei (1987) was estimated with custom scripts on the basis of the R package Hierfstat (Poelstra et al. 2014). The number of usable invariant sites for d_{xy} calculation were identified based on per base pair sequencing coverage of individuals to use only those sites that are covered by at least 5 reads in more than half of the individuals in both populations.

Quantifying similarity of genomic landscapes within and among clades

We used Pearson correlations as a simple means to characterize the degree of co-variation in genome-wide distribution patterns for a given summary statistic. Correlation coefficients were calculated on the basis of homologous windows within and between clades (see above). For intra-population measures (ρ, π) we calculated all possible combinations between two populations (with more than three individuals) *i=1...n* and *j=i+1....n.* For inter-population metrics (F_{ST}, PBS, d_{xy}) we calculated all possible combinations between-population comparisons *I* (e.g. popA vs. popB*), J* (e.g. popC vs. popD) except for flycatcher where F_{ST} was only available for 16 populations comparisons (cf. Burri et al. 2015). This yields a distribution of correlation coefficients for each summary statistic (see also (Vijay et al. 2016)). Significance in co-variation between populations or population comparisons was attributed if more than 95% of the distribution were above zero (significant positive correlation) or below zero (significant negative correlation).

Overlap with centromeres and sub-telomeres

LiftOvers to the zebra finch genome in principle allow associating outlier regions from genome scans (e.g. islands of elevated differentiation) with genomic features such as centromeres or sub-telomeres. This approach works under the assumption of karyotype conservation across large evolutionary timescales (Ellegren 2010). It is conservative in that overlap is only expected if centromere position is conserved between zebra finch and the taxon under consideration. Evolutionary lability of these features, partly expected due to known lineage-specific inversions in zebra finch (Romanov et al. 2014; Kawakami et al. 2014; Hooper and Price 2015) would reduce any real correlation (Type II error), but is unlikely to introduce spurious correlations (Type I error). Twenty-two centromere and 20 sub-telomere positions were obtained for zebra finch from Knief & Forstmeier (2016)□. Regions identified as centromeres were on average ~1Mb long (mean: 960,100 bp; range: 150,000 bp to 5,350,000 bp) while the sub-telomeric regions were shorter (mean:

169800; range: 50,000 bp to 298,700 bp). Some of the sub-telomeric regions and centromeres were located at the extreme ends of the chromosomes and orthologous regions could not be identified in the draft assemblies of the crow, flycatcher and Darwin's finch. These regions are either not assembled in the draft genomes, or synteny could not be unambiguously assigned.

Of the 42 regions that have been identified as centromeres or sub-telomeric regions in zebra finch, orthologous regions could be identified for a subset of 38 in the flycatcher (mean recovery, i.e. mean of the fraction of each of the regions mapped: 0.69), 39 in crow (mean recovery: 0.83) and 25 in the Darwin's Finch genome (mean recovery: 0.55). The relatively low recovery in Darwin's finch is most likely owing to the lower quality of its genome, which is more fragmented than the genomes of flycatcher and, particularly, of crow. The sub-telomeres of chromosome 5, 13 and 21 could be lifted over in neither crow nor flycatcher genomes suggesting a systematic bias for these regions. To reduce the effect of such bias, we not only looked for overlap of outlier peaks (as defined below) with centromeres or sub-telomeres, but also for overlap with increasing distance from the inferred positions of these features in five incremental steps of 10 kb. In the case of random association no relationship would be expected with distance. In the case of genuine association significance of the overlap should decrease with distance.

To relate characteristics of the genomic differentiation landscape to chromosomal features, we proceeded as follows. For each taxon, we chose two independent population comparisons with the highest genome-wide average F_{ST} values. This strategy is owing to the fact that clear 'background peaks' caused by shared linked selection only start crystallising at an advanced level of population divergence (Burri et al. 2015; Vijay et al. 2016). This is theoretically expected and has been shown in crows where an increase in genome-wide F_{ST} is accompanied by an increase in autocorrelation between windows, peak overlap and the degree of co-variation in differentiation landscapes (Vijay et al. 2016). Population pairs used and their corresponding differentiation statistics are shown in **Table S4**. We then used positions along the zebra finch genome to calculate the percent of centromeres and sub-telomeres that overlapped with differentiation outliers (**Table S5**). To check if the percent of overlap we observed was more than that expected by chance, we permuted the positions of centromeres and sub-telomeres within each chromosome 1000 times using the shuffle option in bedtools (Quinlan and Hall 2010) and calculated the percent of overlap that was expected by chance alone. A significant association is inferred at type I error levels of 0.005/ 0.001 if the test

statistic derived from the empirical centromere/sub-telomere distribution exceeded a maximum of 4/0-times by test statistics derived from the permuted distributions.

RESULTS

Co-variation within clades (micro-level)

Previous studies in flycatcher (Burri et al. 2015; Kawakami et al. 2017) and crow (Vijay et al. 2016) have shown that population-scaled recombination rate (ρ), nucleotide diversity (π) and measures of genetic differentiation (F_{ST}, PBS and d_{xy}) were significantly correlated between population (comparisons) within each clade. Extending the population comparison of ρ, π, F_{ST}, PBS and d_{xy} to Darwin's finch complex corroborate the generality of this finding. Genome-wide patterns of these summary statistics summarized in **Fig. 2** and **Table S6** were positively correlated among all populations in each of the three clades. For ρ, correlation coefficients were highest in flycatchers (mean r=0.43), followed by Darwin's finches (r= 0.27) and crows (r=0.19). Nucleotide diversity π showed strongest co-variation in flycatchers (r=0.95), followed by crows (r=0.70) and Darwin's Finches (r=0.49). Correlation of F_{ST} was consistently positive between all population pairs in Darwin's finches (r=0.46), flycatchers (mean r=0.42) and crows (r=0.36). The correlation for PBS was even stronger than F_{ST} (r=0.64 in Darwin's finches, r=0.46 in flycatchers and r=0.42 in crows). d_{xy} showed significantly positive correlations between pairs of populations within each clade with mean correlation coefficients of 0.72, 0.85 and 0.94 in flycatchers, crows and Darwin's finches, respectively. Importantly, d_{xy} was negatively correlated with F_{ST} (mean range r=-0.45 to -0.19). This is predicted by long-term linked selection (acting already in the ancestor) and is opposed to the expectation for divergent selection in the face of gene flow (Nachman and Payseur 2012; Cruickshank and Hahn 2014).

Co-variation across clades (macro-level)

Next, we investigated whether the summary statistics indicative of local N_e used in the intra-clade comparisons also co-varied in syntenic regions between clades. Though effect sizes were lower, correlations were consistently positive for all summary statistics (**Fig. 2B, Table S7**). Mean Pearson's correlation coefficient in the population-scaled recombination rate (ρ) ranged from 0.099 (crow vs. flycatcher) to 0.172 (flycatcher vs. Darwin's finch); for nucleotide diversity (л) from 0.082 (flycatcher vs. Darwin's finch) to 0.271 (crow vs. flycatcher). Patterns of genetic differentiation were also similar between clades with F_{ST} ranging from 0.115 (crow vs. flycatcher)

to 0.163 (crow vs. Darwin's finch) and PBS ranging from 0.185 (crow vs. Darwin's finch) to 0.231 (flycatcher vs Darwin's finch). d_{xy} showed the highest inter-clade correlations ranging from 0.224 (flycatcher vs. Darwin's finch) to 0.342 (crow vs. flycatcher). As in the micro-level comparisons, d_{xy} and F_{ST} were negatively correlated among clades (mean range r=-0.21 to -0.16). The strength of correlation in all of these summary statistics was not systematically associated with divergence time representing 50 million years of independent evolution (**Fig. 2B, Table S7, Fig. S4).**

Overlap with structural genomic features

We next sought to investigate the potential impact of structural genomic features where the effect of linked selection might be particularly pronounced. We evaluated whether regions of highly elevated differentiation were associated with regions of suppressed recombination adjacent to peri-centromeric and sub-telomeric regions as predicted from the location of such regions in zebra finch (karyotype data is not available for both crow and collared flycatcher; **Fig. 3A**). For each clade, we focused on the two most divergent population/species comparisons (Burri et al. 2015; Vijay et al. 2016). In all three clades, the overlap was significantly larger than expected by chance in at least one comparison of each species (percentage of overlap in flycatchers: 58.53% and 60.98%, crows: 21.95% and 31.7%, Darwin's finches: 14.63% and 29.27%) (**Fig. 3B**). When regions next to centromeres and sub-telomeric regions were considered separately, there was a significant association for sub-telomeric regions in all three clades (**Fig. S5**), whereas the association for regions next to centromeres was significant only in flycatcher (**Fig. S6**).

DISCUSSION

In this study we quantified genome-wide patterns of genetic diversity within and between multiple populations for each of three phylogenetically distant avian clades with split times beyond the expected time for complete lineage sorting. We asked the question whether these 'landscapes of genetic diversity' co-varied across micro-evolutionary timescales among populations within clades, and across macro-evolutionary timescales among clades.

As previously reported, genome-wide heterogeneity in genetic variation captured by population genetic statistics reflective of local N_e co-varied among populations within clades. Studies in sunflowers (Renaut et al. 2013, stonechats (Van Doren et al. 2017), crows (Vijay et al. 2016) and flycatchers (Burri et al. 2015) similarly reported that landscapes of variation in genetic diversity

were correlated among populations and closely related species differing in divergence time and the level of gene flow. An explanation for the correlated pattern of diversity, therefore, requires a mechanism universally affecting all populations. Variation in the strength of linked selection mediated by local levels of recombination rate shared among populations has been suggested as a primary force. In flycatchers, for example, where pedigree based recombination rate data is available, linked selection serves an explanation for genomic parallelism among populations and species without the need to invoke population specific adaptation and context-dependent selection in the face of gene flow (Burri et al. 2015).While mutation rate may contribute in shaping genome-wide variation in genetic diversity, linked selection appears to be the dominant mechanism (Dutoit et al. 2017).

The present study adds a macro-evolutionary, comparative axis providing evidence for linked selection at syntenic regions across large phylogenetic distances where any contribution of shared ancestry, gene flow or common environmental factors can be excluded. Summary statistics capturing information on N_e were correlated among clades spanning over 50 millions of years of divergence. The degree of correlation among clades was remarkable considering divergence times of several million generations, gaps in syntenic alignments and the statistical error associated with population genetic estimates from moderate samples sizes. With recombination rate being the key mediator of linked selection, an explanation of genomic parallelism in N_e through linked selection requires conserved recombination landscapes among the clades under investigation. Unlike mammals, relatively stable karyotype in birds (Ellegren 2010) argue for the conservation of recombination landscape; however, the extent of such conservation is not clear, in particular at the level of individual chromosomes. Comparative analysis among chicken, zebra finch and collared flycatcher suggests that intra-chromosomal re-arrangements occurred at non-negligible rates, and that lack of recombination around (macro-)chromosome centres appears to be specific to zebra finch (Kawakami et al. 2014). It is thus not straightforward to predict the degree of co-variation in recombination rates at kb-resolution considered here. The observed correlation in population-scaled recombination rates between clades, however, is consistent with the assumption that overall recombination landscapes are sufficiently similar to mediate common patterns of linked selection. Nevertheless, it has been suggested that recombination rate could slightly change even within clades in birds (Kawakami et al. 2017), indicating that genetic diversity and differentiation could evolve in a species or clade-specific manner. It should further be noted that mutation rate variation ($\mu \sim N_e$;

$r \sim N_e \mu$) could also contribute to the correlation. However, compared to the effect of recombination rate its effect on genome-wide variation of genetic diversity seems minor (Cutter and Payseur 2013; Dutoit et al. 2017).

The magnitude of correlations of all summary statistics was not related to divergence time (**Fig. S4**) with sometimes noticeably higher correlation coefficients for the phylogenetically older flycatcher-crow comparison, than for the younger flycatcher-finch comparison (**Table S7**). This suggests that the strength of co-variation may be underestimated by factors such as genome quality, population sampling and/or differences in the degree of rearrangements between clades. Due to these limitations, a direct comparison of effect sizes between intra- and inter-clade comparisons which would allow the separation of population-specific selection from selection shared across all clades under consideration is at present not possible. However, substantial co-variation among clades indicates that genomic regions with properties amenable to linked selection reducing N_e remained stable across millions of years of evolution. The observation that d_{xy} was generally reduced in areas of high relative differentiation (F_{ST}, PBS) both within and across clades points towards a selective process continuously purging diversity and reducing effective population size (Cruickshank and Hahn 2014). Van Doren and co-authors (2017) also reported co-variation in $F_{ST,}$ d_{xy} and π across the shorter evolutionary distance between flycatchers and stonechat, and similarly concluded that linked selection continuously erodes local genetic diversity possibly before the divergence of these species.

Linked selection can occur in the form of background selection (Charlesworth 1994) or recurrent hitch-hiking dynamics by selective sweeps (Smith and Haigh 1974). Consistent with both types of selection, recent population genetic studies of flycatchers and crows suggest that diversity and differentiation landscapes were associated with variation in recombination rate and gene density (as a proxy for the target of selection) within clades (Burri et al. 2015; Vijay et al. 2016). In species with moderate effective population sizes, beneficial mutations are expected to be limited, and the distribution of fitness effects are likely to differ between species (Eyre-Walker and Keightley 2007). Parallel positive selection forming the basis of adaptation or divergent selection affecting the same genomic regions in different clades is thus expected to be rare. Background selection on the other hand appears to be less limited by mutational input, assuming that the vast majority of new mutations are deleterious. Given its long-term effects it will also be only slightly affected by the

transitory population-specific demographic change (Ewing and Jensen 2016; Beissinger et al. 2016; Coop 2016). Based on model-based coalescent simulation, Corbett-Detig and co-workers (2015) suggested that for species with low/moderate population sizes (including flycatchers) background selection would prevail over hitch-hiking in relative importance (but see Coop (2016) and Munch et al. (2016)). Importantly, linked selection based on either background selection or selective sweeps will reduce ancestral genetic variation and consequently generate shared patterns of reduced genetic diversity in low recombination regions. The observed negative correlation between F_{ST} and d_{xy} is consistent with predictions of linked selection of both background and positive selection reducing not only population-specific, but ancestral genetic variation. Yet, it cannot fully be excluded that loci directly governing population specific adaptation or promoting population divergence can emerge in parallel among clades. Such an explanation would, however, need to invoke continuous and frequent occurrences of selective sweeps reducing genetic variation at syntenic regions between clades. The inclusion of more species from larger evolutionary distances with distinct biogeographic histories will help to further resolve the relative contribution of factors influencing local genetic diversity.

In all clades under investigation, we found evidence for reduced diversity and elevated differentiation at candidate (peri-)centromeric regions. A similar association was suggested for mouse (Carneiro et al. 2009), Swainson's thrushes (Delmore et al. 2015) and stickleback fish (Roesti et al. 2013). These studies are consistent with the idea that strongly reduced recombination rate in the vicinity of centromeres will most strongly be affected by linked selection. However, centromeric positions in crow, flycatcher and Darwin's finch were approximated relative to centromeres in zebra finch. Zebra finch is known for its many lineage-specific inversions (Kawakami et al. 2014; Weissensteiner et al. 2017) which may have reduced the association of genetic differentiation with the predicted centromere locations in the target species. Recent work in crows, however, corroborates an impact of independently predicted, putative (peri-)centromeric regions on population recombination, genetic diversity and differentiation (Weissensteiner et al. 2017). In addition to putative centromeric regions, we found evidence for an association of sub-telomeric regions with variation in genetic diversity. Yet, sub-telomeric regions are not necessarily characterized by low recombination in birds (Backström et al. 2010; Kawakami et al. 2014) which is consistent with an explanation invoking recurrent positive selection rather than background selection reducing local N_e. However, in other systems it has been shown that sub-telomeric regions

experience low recombination rates, similar to centromeres (Roesti et al. 2013). Further evaluation of this hypothesis will require fine-scale recombination rate estimates across all clades.

In conclusion, we advocate the use of comparative, phylogenetic approaches to shed light on population-level processes introducing heterogeneity in patterns of diversity, differentiation, and divergence along the genome. Most insight will be gained in taxa with high-quality, chromosome level genome assemblies with correct placement of centromeric and sub-telomeric regions. Independent estimates of mutation and recombination rates are further crucial to assess the genomic stability of these central processes across evolutionary timescales. On the bioinformatic side, unbiased methods for translating orthologous genomic coordinates among a large number of distantly related species are required.

ACKNOWLEDGMENTS

Funding for this study was provided by the Swedish Research Council (grant number 621-2010-5553 to J.W., 2014-6325 to T.K. and 2013-08721 to H.E.), Marie Sklodowska Curie Actions (grant number 600398 to T.K.), the European Research Council (grant number ERCStG-336536 to J.W.), the Knut and Alice Wallenberg Foundation (to H.E. and J. W.) and the Swiss National Science Foundation (grants number PBLAP3-134299 and PBLAP3_140171 to R.B.). We are grateful for the access to the computational infrastructure provided by the UPPMAX Next-Generation Sequencing Cluster and Storage (UPPNEX) project, funded by the Knut and Alice Wallenberg Foundation and the Swedish National Infrastructure for Computing. We would like to thank Leif Andersson and his group for providing access to the genotype data from Lamichhaney et al. (2015). We are also grateful to Claire Peart for valuable input on the manuscript.

Data Accessibility

Raw data forming the basis for this study are publicly available at PRJNA192205 & PRJEB9057 (Crows), PRJEB2984 (Flycatchers), PRJNA301892 (Darwin's Finches).

Author Contributions

NV and JW conceived the study, NV conducted all bioinformatic analyses with help from MW. RB, TK and HE provided population genetic summary statistics for the flycatcher. NV and JW wrote the manuscript with input from all other authors.

REFERENCES

Abbott RD, Albach S, Ansell et al. (2013). Hybridization and Speciation. *Journal of Evolutionary Biology,* 26, 229–46.

Auton A, Adi FA, Pfeifer S et al. (2012). A Fine-Scale Chimpanzee Genetic Map from Population Sequencing. *Science,* 336, 193–98.

Backström N, Wolfgang F, Schielzeth H et al. (2010). The Recombination Landscape of the Zebra Finch *Taeniopygia Guttata* Genome. *Genome Research,* 20, 485–95.

Bank C, Ewing GB, Ferrer-Admettla A et al. (2014). Thinking Too Positive? Revisiting Current Methods of Population Genetic Selection Inference. *Trends in Genetics*, 30, 540–46.

Barton NH. (1983). Multilocus Clines. *Evolution,* 37, 454–71.

Barton NH and Bengtsson BO (1986) The Barrier to Genetic Exchange between Hybridising Populations. *Heredity,* 57, 357–76.

Beissinger TM, Li W, Kate C et al. (2016). Recent Demography Drives Changes in Linked Selection across the Maize Genome. *Nature Plants,* 2, 16084.

Brommer JE, Gustafsson L, Pietiäinen H et al. (2004). Single-Generation Estimates of Individual Fitness as Proxies for Long-Term Genetic Contribution. *The American Naturalist,* 163, 505–517.

Burri R, Nater A, Kawakami T 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 Research,* 25, 1656–65.

Carneiro M, Frank WA, Sandra A et al. (2014). The Genomic Architecture of Population Divergence between Subspecies of the European Rabbit. *PLOS Genetics,* 10, e1003519.

Carneiro M, Nuno F and Nachman MW (2009). Recombination and Speciation: Loci near Centromeres Are More Differentiated than Loci near Telomeres between Subspecies of the European Rabbit (Oryctolagus Cuniculus). *Genetics,* 181, 593–606.

Chan AH, Jenkins PA and Song YS. (2012). Genome-Wide Fine-Scale Recombination Rate Variation in Drosophila Melanogaster. *PLOS Genetics,* 8, e1003090.

Charlesworth, B. (1994). The Effect of Background Selection against Deleterious Mutations on Weakly Selected, Linked Variants. *Genetical Research,* 63, 213–27.

Charlesworth, B. (1998). Measures of Divergence between Populations and the Effect of Forces That Reduce Variability. *Molecular Biology and Evolution,* 15, 538–43.

Charlesworth B, Morgan MT and Charlesworth D (1993). The Effect of Deleterious Mutations on

Neutral Molecular Variation. *Genetics*, 134, 1289–1303.

Coop G (2016). Does Linked Selection Explain the Narrow Range of Genetic Diversity across Species? *bioRxiv*, March, 042598.

Corbett-Detig RB, Hartl DL and Sackton TB (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biol*, 13, e1002112.

Cruickshank TE and Hahn MW (2014). Reanalysis Suggests That Genomic Islands of Speciation Are due to Reduced Diversity, Not Reduced Gene Flow. *Molecular Ecology*, 23, 3133–57.

Cutter AD and Payseur BA (2013). Genomic Signatures of Selection at Linked Sites: Unifying the Disparity among Species. *Nature Reviews. Genetics*, 14, 262–74.

Delmore KE, Hübner S, Kane NC et al. (2015). Genomic Analysis of a Migratory Divide Reveals Candidate Genes for Migration and Implicates Selective Sweeps in Generating Islands of Differentiation. *Molecular Ecology*, 24, 1873–88.

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, 491–98.

Dettman JR, Sirjusingh C, Kohn LM et al. (2007). Incipient Speciation by Divergent Adaptation and Antagonistic Epistasis in Yeast. *Nature*, 447, 585–88.

Dutoit L, Vijay N, Mugal CF et al. (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.

Ellegren H (2010). Evolutionary Stasis: The Stable Chromosomes of Birds. *Trends in Ecology & Evolution*, 25, 283–91.

Ellegren H and Galtier N (2016). Determinants of Genetic Diversity. *Nature Reviews Genetics*, 17, 422–33.

Ellegren H, Smeds L, Burri R et al. (2012). The Genomic Landscape of Species Divergence in Ficedula Flycatchers. *Nature*, 491, 756–60.

Ericson PGP, Zuccon D, Ohlson JI et al. (2006). Higher-Level Phylogeny and Morphological Evolution of Tyrant Flycatchers, Cotingas, Manakins, and Their Allies (Aves: Tyrannida). *Molecular Phylogenetics and Evolution*, 40, 471–83.

Ewing GB and Jeffrey DJ. (2016). The Consequences of Not Accounting for Background Selection in Demographic Inference. *Molecular Ecology*, 25, 135–41.

Eyre-Walker A and Keightley PD. (2007). The Distribution of Fitness Effects of New Mutations. *Nature Reviews Genetics*, 8, 610–18.

Gompert Z and Buerkle CA. (2011). Bayesian Estimation of Genomic Clines. *Molecular Ecology,* 20, 2111–27.

Goudet J (2005). Hierfstat, a Package for R to Compute and Test Hierarchical F-Statistics. *Molecular Ecology Notes,* 5, 184–86.

Grant PR and Grant BR (1992). Hybridization of Bird Species. *Science,* 256, 193–97.

Harris RS. (2007). Improved Pairwise Alignment of Genomic DNA. Phd thesis, Pennsylvania State University.

Hinrichs AS, Karolchik D, Baertsch R et al. (2006). The UCSC Genome Browser Database: Update (2006). *Nucleic Acids Research,* 34, D590-598.

Hodgkinson A and Eyre-Walker A (2011). Variation in the Mutation Rate across Mammalian Genomes. *Nature Reviews Genetics,* 12, 756–66.

Hooper DM and Price TD (2015). Rates of Karyotypic Evolution in Estrildid Finches Differ between Island and Continental Clades. *Evolution,* 69, 890–903.

Hubbard TD, Birney BE, Cameron G et al. (2002). The Ensembl Genome Database Project. *Nucleic Acids Research,* 30, 38–41.

Hudson RR and Coyne JA (2002). Mathematical Consequences of the Genealogical Species Concept. *Evolution,* 56, 1557–65.

Jarvis ED, Mirarab S, Aberer AJ et al. (2014). Whole-Genome Analyses Resolve Early Branches in the Tree of Life of Modern Birds. *Science,* 346, 1320–31.

Jetz W, Thomas GH, Joy JB et al. (2012). The Global Diversity of Birds in Space and Time. *Nature,* 491, 444–48.

Jetz W, Gavin H. Jeffrey TB et al. (2014). Global Distribution and Conservation of Evolutionary Distinctness in Birds. *Current Biology,* 24, 919–30.

Knud Andreas J, Fabre PH, Kennedy JD et al. (2016). A Supermatrix Phylogeny of Corvoid Passerine Birds (Aves: Corvides). *Molecular Phylogenetics and Evolution* 94, Part A: 87–94.

Kawakami T, Mugal CF, Suh A et al. (2017) Whole-Genome Patterns of Linkage Disequilibrium across Flycatcher Populations Clarify the Causes and Consequences of Fine-Scale Recombination Rate Variation in Birds." In review

Kawakami T, Smeds L, Backström N et al. (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. *Molecular Ecology,* 23, 4035–

58.

Knief U and Forstmeier W (2016). Mapping Centromeres of Microchromosomes in the Zebra Finch (Taeniopygia Guttata) Using Half-Tetrad Analysis. *Chromosoma*, 125, 757–68.

Kronforst MR and Papa R (2015). The Functional Basis of Wing Patterning in Heliconius Butterflies: The Molecules Behind Mimicry. *Genetics*, 200, 1–19.

Kuhn RM, Karolchik D, Zweig AS et al. (2007). The UCSC Genome Browser Database: Update 2007. *Nucleic Acids Research*, 35, D668-673.

Lamichhaney S, Berglund J, Sällman Almén M et al. (2015). Evolution of Darwin's Finches and Their Beaks Revealed by Genome Sequencing. *Nature*, 518, 371–75.

Lewontin RC and Krakauer J. (1973). Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. *Genetics*, 74, 175–95.

Mallet J, Beltrán M, Neukirchen W et al. (2007). Natural Hybridization in Heliconiine Butterflies: The Species Boundary as a Continuum. *BMC Evolutionary Biology*, 7, 28.

Munch K, Kiwoong N, Schierup MH et al. (2016). Selective Sweeps across Twenty Millions Years of Primate Evolution. *Molecular Biology and Evolution*, 33, 3065–74.

Nachman MW and Payseur BA (2012). Recombination Rate Variation and Speciation: Theoretical Predictions and Empirical Results from Rabbits and Mice. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 367, 409–21.

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.

Nei M (1987). Molecular Evolutionary Genetics. Columbia University Press. (equation 10.20)

Nosil P and Feder JL (2013). Genome Evolution and Speciation: Toward Quantitative Descriptions of Pattern and Process. *Evolution*, 67, 2461–67.

Poelstra JW, Vijay N, Bossu CM et al. (2014). The Genomic Landscape Underlying Phenotypic Integrity in the Face of Gene Flow in Crows. *Science*, 344, 1410–14.

Poelstra JW, Vijay N, Hoeppner MP et al. (2015). Transcriptomics of Colour Patterning and Coloration Shifts in Crows. *Molecular Ecology*, 24, 4617–28.

Powell THQ, Glen RH, Mason OM et al., (2013). Genetic Divergence Along the Speciation Continuum: The Transition from Host Race to Species in Rhagoletis (Diptera: Tephritidae). *Evolution*, 67, 2561–76.

Prum RO, Berv JS, Dornburg A et al. (2015). A Comprehensive Phylogeny of Birds (Aves) Using

Targeted next-Generation DNA Sequencing. *Nature*, 526, 569–73.

Puzey JR, Willis JH and Kelly JK. (2017). Population Structure and Local Selection Yield High Genomic Variation in Mimulus Guttatus. *Molecular Ecology*, 26, 519–35.

Quinlan AR and Hall IM (2010). BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics*, 26, 841–42.

Rands CM, Darling A, Fujita M et al. (2013). Insights into the Evolution of Darwin's Finches from Comparative Analysis of the Geospiza Magnirostris Genome Sequence. *BMC Genomics*, 14, 95.

Renaut S, Grassa CJ, Yeaman S et al. (2013). Genomic Islands of Divergence Are Not Affected by Geography of Speciation in Sunflowers. *Nature Communications*, 4, 1827.

Roesti M, Hendry AP, Salzburger W et al. (2012). Genome Divergence during Evolutionary Diversification as Revealed in Replicate Lake–stream Stickleback Population Pairs. *Molecular Ecology*, 21, 2852–2862.

Roesti M, Kueng B, Moser D et al. (2015). The Genomics of Ecological Vicariance in Threespine Stickleback Fish. *Nature Communications*, 6, 8767.

Roesti M, Moser S and Berner D (2013). Recombination in the Threespine Stickleback Genome--Patterns and Consequences. *Molecular Ecology*, 22, 3014–27.

Romanov MN, Farré M, Lithgow PE et al. (2014). Reconstruction of Gross Avian Genome Structure, Organization and Evolution Suggests That the Chicken Lineage Most Closely Resembles the Dinosaur Avian Ancestor. *BMC Genomics*, 15, 1060.

Scheet P and Stephens M (2006). A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *American Journal of Human Genetics*, 78, 629–44.

Seehausen O, Butlin RK, Keller I et al. (2014). Genomics and the Origin of Species. *Nature Reviews Genetics*, 15, 176–92.

Shafer ABA, Peart CR, Tusso S et al. (2016). Bioinformatic Processing of RAD-Seq Data Dramatically Impacts Downstream Population Genetic Inference. *Methods in Ecology and Evolution*, online early.

Singhal S, Leffler EM, Sannareddy K et al. (2015). Stable Recombination Hotspots in Birds. *Science*, 350, 928–32.

Slotte T (2014). The Impact of Linked Selection on Plant Genomic Variation. *Briefings in Functional Genomics*, 13, 268–75.

Smith JM and Haigh J (1974). The Hitch-Hiking Effect of a Favourable Gene. *Genetical Research*, 23, 23–35.

Smith T and Eyre-Walker A (2017). Large Scale Variation in the Rate of de Novo Mutation in Humans and Its Relationship to Divergence and Diversity. *bioRxiv*, February, 110452.

Soria-Carrasco V, Gompert Z, Comeault AA et al. (2014). Stick Insect Genomes Reveal Natural Selection's Role in Parallel Speciation. *Science*, 344, 738–42.

Stephan W (2010). Genetic Hitchhiking versus Background Selection: The Controversy and Its Implications. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 365, 1245–53.

Strasburg JL, Sherman NA, Wright KM et al. (2012). What Can Patterns of Differentiation across Plant Genomes Tell Us about Adaptation and Speciation? *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 367, 364–73.

Tang H, Li J and Krishnakumar V. (2015). Jcvi: JCVI Utility Libraries.

Tine M, Kuhl H, Gagnaire PA et al. (2014). European Sea Bass Genome and Its Variation Provide Insights into Adaptation to Euryhalinity and Speciation. *Nature Communications*, 5, 5770.

Van Doren BM, Campagna L, Helm B et al. (2017). Correlated Patterns of Genetic Diversity and Differentiation across an Avian Family. *Molecular Ecology*, In Press

Vijay N, Bossu CM, Poelstra JW et al. (2016). Evolution of Heterogeneous Genome Differentiation across Multiple Contact Zones in a Crow Species Complex. *Nature Communications*, 7, 13195.

Vijay N, Poelstra JW, Künstner A et al.(2013). Challenges and Strategies in Transcriptome Assembly and Differential Gene Expression Quantification. A Comprehensive in Silico Assessment of RNA-Seq Experiments. *Molecular Ecology*, 22, 620–34.

Weber CC, Boussau B, Romiguier J et al. (2014). Evidence for GC-Biased Gene Conversion as a Driver of between-Lineage Differences in Avian Base Composition. *Genome Biology*, 15: 549.

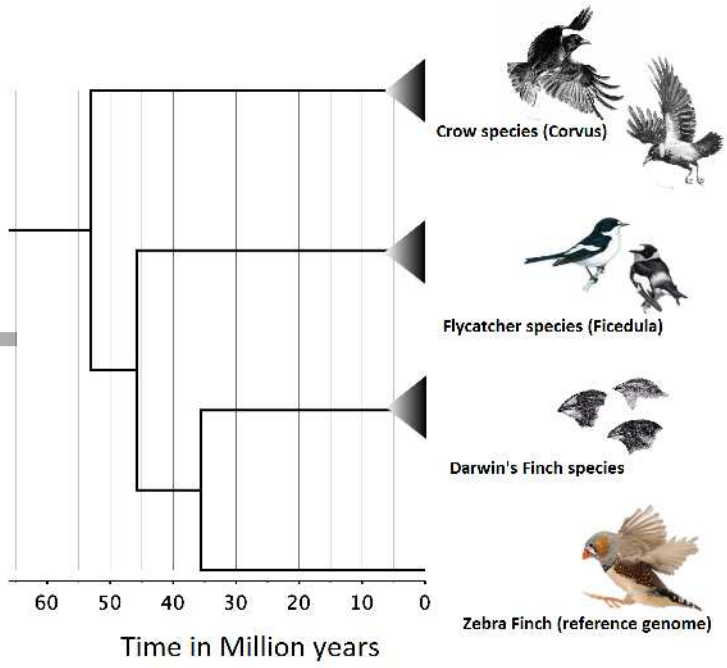
Weissensteiner M, Pang A, Bunikis I et al. (2017). Combination of Short-Read, Long-Read and Optical Mapping Assemblies Reveals Presumably Heterochromatic Tandem Repeat Arrays with Population Genetic Implications. *Genome Research*, 27, 697-708

Wolf JBW, Bayer T, Haubold B et al. (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. *Molecular Ecology*, 19, 162–75.

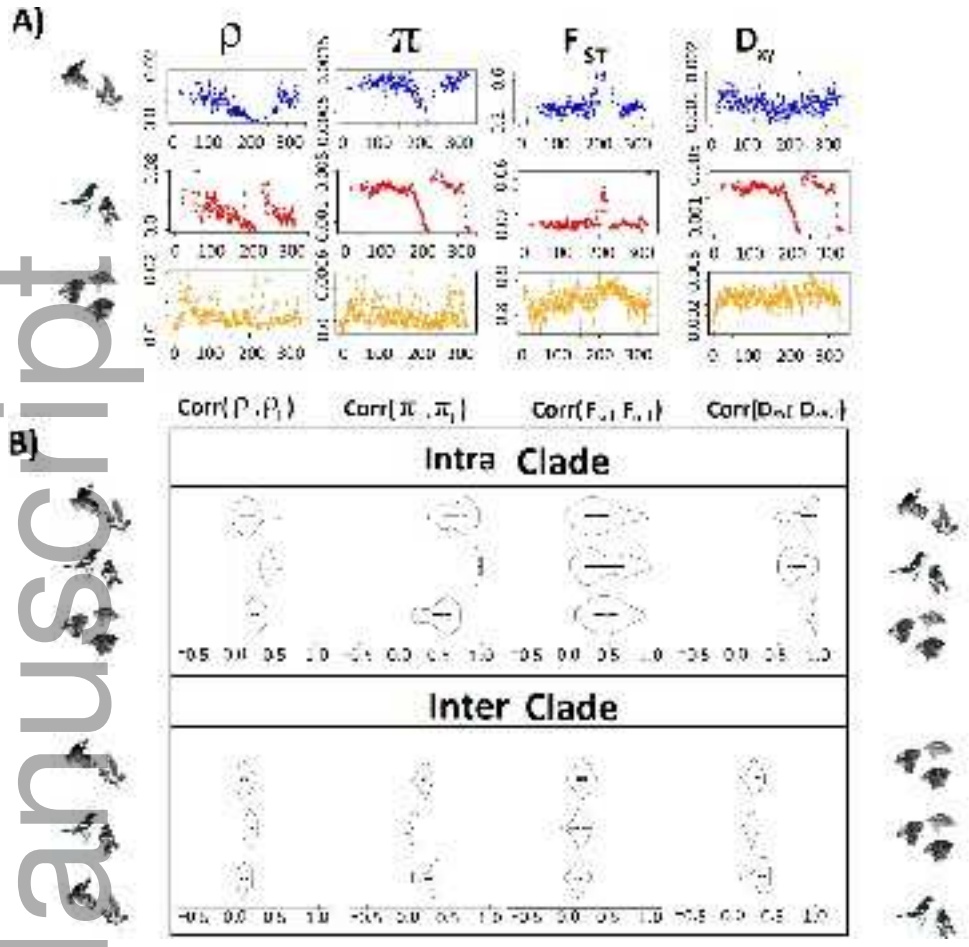
Wolf JBW and Ellegren H (2017). Making Sense of Genomic Islands of Differentiation in Light of Speciation. *Nature Reviews Genetics*, 18, 87–100.

Wolf JBW, Lindell J and Backström N. (2010). Speciation Genetics: Current Status and Evolving Approaches. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365, 1717–33.

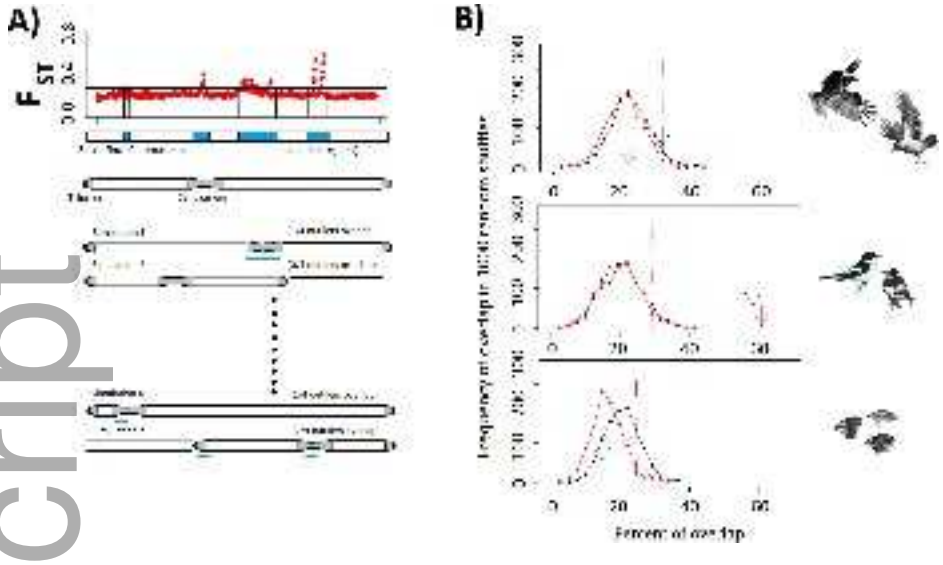
Wu CI. (2001). The Genic View of the Process of Speciation. *Journal of Evolutionary Biology*, 14, 851–65.



mec_14195_f1.tiff



mec_14195_f2.tiff



mec_14195_f3.tiff