How linked selection shapes the diversity landscape in

_Ficedula_ flycatchers

Agnes Rettelbach*, Alexander Nater*,† & Hans Ellegren*

* Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden

† Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, D-78457 Konstanz, Germany
Linked selection in flycatchers

Keywords
Diversity
Background selection
Selective sweeps
Flycatcher

Corresponding author
Agnes Rettelbach
Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, SE-752 36 Uppsala, Sweden
agnes.rettelbach@technikum-wien.at
ABSTRACT

There is an increasing awareness that selection affecting linked neutral sites strongly influences on how diversity is distributed across the genome. In particular, linked selection is likely involved in the formation of heterogenous landscapes of genetic diversity, including genomic regions with locally reduced effective population size that manifests as dips in diversity and “islands” of differentiation between closely related populations or species. Linked selection can be in the form of background selection or selective sweeps and a long-standing question in population genetics has been to unveil the relative importance of these processes. Here we analysed the theoretically expected reduction of diversity caused by linked selection in the collared flycatcher (*Ficedula albicollis*) genome and compared this with population genomic data on the distribution of diversity across the flycatcher genome. By incorporating data on recombination rate variation and the density of target sites for selection (including both protein-coding genes and conserved non-coding elements), we found that background selection can explain most of the observed base line variation in genetic diversity. However, positive selection was necessary to explain the pronounced local diversity dips in the collared flycatcher genome. We confirmed our analytical findings by comprehensive simulations. Our study therefore demonstrates that even though both background selection and selective sweeps contribute to the heterogeneous diversity landscape seen in this avian system, they play different roles in shaping it.
INTRODUCTION

Natural selection reduces genetic diversity at neutral sites mainly by two key concepts of population genetics: hitchhiking (Maynard Smith and Haigh, 1974; Kaplan et al., 1989, Fay & Wu 2000) and background selection (BGS, Charlesworth et al., 1993, Hudson & Kaplan, 1995, Nordborg et al. 1996). The former process means that positive selection for an advantageous variant also selects for the genetic background (haplotype) on which the beneficial allele resides. The latter process implies that purifying selection against recurring deleterious mutations decreases the diversity at linked neutral sites.

An advantageous allele is expected to quickly increase in frequency and reach fixation in the population, thereby being a potent diversity-reducing force in large genomic regions linked to (hitchhiking with) the site under positive selection. Removal of deleterious variants is typically a slower process if the fitness effects are small and/or the effective population size is low (Crow & Kimura 1970). During the time a disadvantageous allele still segregates in the population, recombination breaks up linkage to nearby variants and thereby narrows the region in which genetic diversity will become reduced by purifying selection. Importantly, background selection only removes haplotypes on which deleterious alleles reside, leaving variants carried by other haplotypes free to segregate. Single episodes of purifying selection therefore do not have the same diversity-reducing effect as positive selection (e.g. Stephan 2010). However, since the distribution of fitness effects is typically strongly biased towards deleterious mutations (Eyre-Walker & Keightley 2007), the combined effect of selection against many deleterious mutations means that background selection is likely to have an as strong, or even stronger, effect on genetic diversity as selective sweeps. Elucidating the relative importance of these two types of linked selection, and under which circumstances one or the other process dominates, has been a challenging question (Charlesworth, 1996; Kim
and Stephan, 2000; McVicker et al., 2009; Comeron, 2014; Elyashiv et al., 2016). The feasibility of obtaining large-scale population genomic data together with data on recombination rate variation and detailed genome annotation now offer exciting possibilities to understand the underlying processes behind linked selection.

If linked selection is prevalent and acts genome-wide, this should be visible as correlations between diversity and factors affecting the extent of linked selection. First, genomic regions with high recombination rate are expected to experience less linked selection since, as indicated above, recombination decouples linked loci and restricts the area of effect of a selected mutation (e.g. Kaplan et al. 1989, Nordborg et al. 1996). Diversity should thus be positively correlated with recombination rate (see for example Mackay et al. 2012). Second, regions with a high density of potential targets for selection are expected to experience more linked selection simply because selection occurs more often in such regions. Diversity should thus be negatively correlated with the density of target sites for selection. Indeed, such correlations have been found in several organisms (e.g., Begun and Aquadro, 1992; Nachman, 2001; Tenaillon et al 2001, Nordborg et al. 2005, Cutter and Payseur 2013, Burri et al., 2015). Third, mutation rate variation influences the number of newly arising mutations under purifying selection and thus also the local strength of linked selection.

Given these relationships and with access to appropriate data, it should be possible to predict genetic diversity under varying influence of background selection and selective sweeps (Charlesworth, 1996; McVicker et al., 2009; Comeron, 2014; Elyashiv et al., 2016; Comeron, 2017).

Linked selection is of particular relevance in the context of speciation (Seehausen et al. 2014; Wolf & Ellegren 2017). During speciation with gene flow, divergent selection at loci
underlying ecologically relevant traits or genetic incompatibilities will hinder gene flow in the vicinity of such loci. In turn, this will lead to localized signals of increased differentiation and reduced diversity, and the formation of so-called speciation islands (Nosil, 2008; Nosil et al., 2008; Feder and Nosil, 2010). However, it has recently been recognised that linked selection can produce similar patterns even in the absence of gene flow (Charlesworth 1998, Noor and Bennett, 2009; Turner and Hahn, 2010; Cruickshank and Hahn, 2014; Burri et al., 2015). Linked selection will locally reduce the effective population size ($N_e$) and thereby not only reduce diversity but also lead to elevated measures of relative differentiation between diverging lineages. If the extent of linked selection varies across the genome, so will also the extent of differentiation. The significance of genomic islands of differentiation in speciation is thus disputed (see Ravinet et al. 2017 for a review).

*Ficedula* flycatchers represent a well-studied model system for speciation research (Lundberg and Alatalo 1992) and constitute one of the most prominent examples of species with distinct differentiation islands (which fully coincide with the location of dips in genetic diversity) spread across the genome (Ellegren et al. 2012, Burri et al. 2015). Relative differentiation ($F_{ST}$) between these recently (<1-2 million years; Nater et al. 2015) diverged species is negatively correlated with recombination rate, indicating a role of linked selection in generating heterogeneous diversity/differentiation landscapes (Burri et al., 2015). Here, we aim to disentangle the forces responsible for shaping the genomic diversity landscape of *Ficedula* flycatchers by modelling the expected impact of linked selection under various scenarios of positive and purifying selection. We compare these results with extensive genome-wide resequencing data and augment the analyses with simulations. We benefit from the access to recombination rate data obtained from linkage analysis (Kawakami et al. 2014) and genome annotation including not only protein-coding genes but also conserved non-
coding elements (Craig et al. 2017). While we find that background selection is generally sufficient to explain the base line levels of genetic diversity in the flycatcher genome, our study suggests that selective sweeps are necessary to generate the most pronounced diversity dips.

METHODS

Genomic and population genomic data

Estimates of nucleotide diversity (\(\pi\)) in non-overlapping 200-kb windows of the collared flycatcher (\(Ficedula albicollis\)), obtained from whole-genome re-sequencing of extensive population samples (79 individuals from four populations), were taken from Burri et al. (2015). To estimate SFS-based statistics, an approach was used that integrates over genotype likelihoods as implemented in the software ANGSD (Korneliussen et al. 2014). This method accounts for genotype uncertainty and considerably improves local estimates of \(\pi\) for low and medium coverage data. Additionally, the sequence data was carefully filtered to avoid biases caused by poorly aligned reads or low sequencing coverage. Briefly, repetitive regions in the reference genome were masked with a custom flycatcher-specific repeat library. Only sites fulfilling the following criteria were considered for calculation of \(\pi\): minimum mapping quality of 1, minimum base quality of 20, site coverage across all individuals less than five standard deviations above the mean coverage, and a minimum read coverage of 5x per individual with a minimum of 10 callable genotypes per population.

We used a discrete time hidden Markov model implemented in the ‘HiddenMarkov’ R package (https://CRAN.R-project.org/package=HiddenMarkov) to classify the window-based \(\pi\) estimates into background regions and diversity valleys. For the observed process dependent on the two hidden states (background and diversity valley), we assumed two normal distributions with standard deviations fixed to the standard deviation of the empirical distribution of \(\pi\) values. We then optimized the means of the distributions for the two hidden
states with the Baum-Welch algorithm (Baum et al. 1970) with a maximum number of 1, 000 iterations. After parameter estimation, we used the Viterbi algorithm to find the most likely sequence of hidden states and identify genomic regions with a predicted diversity valley state. Collared flycatcher recombination rate estimates in cM/Mb for 200-kb windows were taken from Kawakami et al. (2014) and were based on a high-density genetic map containing 4,302 markers. To generate this map, 609 collared flycatcher individuals from a four-generation pedigree were genotyped with a custom-designed SNP array for 37,262 polymorphic loci. Using recombination rate estimates from a pedigree-based linkage analysis instead of higher-resolution LD-based recombination maps (e.g., Kawakami et al 2017) has the advantage that they, in contrast to the latter, are not affected by selection (i.e. it allows direct estimation of recombination rate [r]). Coordinates for conserved regions in the *F. albicollis* genome, including non-coding regions, were taken from Craig et al. (2017). Exon information was obtained from the Ensembl annotation of the collared flycatcher genome assembly version FicAlb 1.4.

**Expected reduction of π due to linked selection**

We compared three different models to calculate the expected deviation from neutral diversity π₀ due to selection. Model 1 was with only background selection, model 2 was model 1 plus recent sweeps, and model 3 was model 2 plus ancient sweeps. For an overview over model parameters, see Table 1.

Under neutrality, nucleotide diversity \( \pi = 4N_e \mu \), where \( N_e \) is the effective population size and \( \mu \) the mutation rate. Mutations at neutral sites are assumed to have no effect on fitness. We used equations from Hudson and Kaplan (1995) and Nordborg et al. (1996) to calculate the
expected deviation from neutral diversity due to background selection, $B = \pi/\pi_0$ (McVicker et al., 2009). $B$ at a focal neutral site $x$ can be approximated by

$$B(x) = \exp \left( -\sum_{i=1}^{n} \frac{u_d \cdot s_d}{(s_d + r(x,i))^2} \right)$$  \hspace{1cm} (1)$$

where $\pi_0$ is without selection and with free recombination, the sum is over all selected sites, $u_d$ is the deleterious mutation rate per site, which is the neutral mutation rate times the proportion of selection targets. $s_d$ is the selection coefficient against heterozygotes, and $r(x,i)$ is the recombination probability between the focal neutral site and the selected site $i$. For $s_d$, the distribution of fitness effects of new mutations at non-synonymous sites estimated for collared flycatcher was used, but with distinct values instead of intervals (13% $Ns=-1$, 9% $Ns=-10$, 14% $Ns=-100$, 64% $Ns=-500$, compare to Bolivar et al. (2018)). As it would be too computationally intensive to calculate $B$ for every site along a chromosome, we implemented a window-based approach with chromosomes divided into non-overlapping 1-kb windows. To calculate genetic distances between two sites, all sites (neutral and selected) were assumed to be located at the midpoint of their respective window. $B$ for each window was thus obtained by considering a neutral site in the midpoint and calculating the influence of selected sites in all linked windows. The smaller the window size, the more exact is this approach. In order to compare the model outcome with $\pi$ from the data, which was available in 200-kb windows to match in turn the recombination rate data, mean $B$ for the respective 200-kb windows was calculated from the mean of the 1-kb windows.

As a measure of the density of sites under purifying selection we used the number of conserved sites (Craig et al. 2017) per window and assumed that all mutations occurring at
these sites were deleterious. As $\pi$ estimates from the data include sites under selection, the model has to include direct selection as well. For each window, the allele frequencies at conserved sites were assumed to be in mutation-selection balance. The expected frequency for additive deleterious mutations is $p = \frac{\mu}{0.5s_d}$ (Crow et al. 1970) and for each conserved site in the window $\pi = p(1 - p)$. Using the distribution of fitness effects of new mutations estimated for collared flycatcher (Bolivar et al. 2018) and a mutation rate of $4.6 \times 10^{-9}$ (Smeds et al., 2016), we estimated mean $\pi=0.0006$ for sites under purifying selection. Note that this is simply a model assumption.

Following Elyashiv et al. (2016), the effect of positive selection can be incorporated in the model. The rate of coalescence at a neutral position $x$ due to a selective sweep at a selected position $j$ is

$$S(x) = \frac{1}{T} \alpha \sum_{j=1}^{k} \exp[-r(x,j)\tau]s_b,$$

(2)

where $T$ is the length of the lineage for which substitutions are considered, $\alpha$ the fraction of substitutions that are beneficial, $s_b$ is the selection coefficient, and $\tau$ is the expected time to fixation of a positively selected mutation, which depends on $N_e$ and $s_b$ (see Elyashiv et al., 2016). As candidates for sites that have undergone a selective sweep, we used non-synonymous substitutions fixed in the collared flycatcher since the split from pied flycatcher ($F. hypoleuca$) for recent sweeps or from red-breasted flycatcher ($F. parva$, with $F. hyperythra$ as outgroup) for ancient sweeps. We set the divergence time between collared flycatcher and pied flycatcher to 320,000 generations, and between collared flycatcher and red-breasted flycatcher to 2,500,000 generations (Nater et al., 2015). We used an estimate of the long-term
Ne in collared flycatcher of 450,000 (Nater et al., 2015). In general, we used \( s_b = 0.1 \) and \( \alpha = 1 \) for the recent sweeps, thus treating all non-synonymous substitutions as strong sweeps, as a contrast to the BGS only model. For the ancient sweeps we generally used \( \alpha = 0.2 \) (~0.18 estimated for flycatcher, Bolivar et al. 2018). However, we also investigated the effect on the diversity landscape when using different values for \( \alpha \) and \( s_b \).

Following Elyashiv et al. (2016), total \( \pi \) is

\[
\pi(x) = \frac{2\mu}{2\mu + \frac{1}{(2N_e \theta(x)) + S(x)}}.
\]

To calculate the background selection parameter \( B \) from the data \( (B_d) \), we assumed a neutral \( \pi_0 = 0.0048 \) (genome-wide average is 0.0041). This was obtained by averaging over all 200-kb windows with \( \text{dcs} < 0.05 \) and recombination rate >3 cM/Mb, which should represent the genomic regions least influenced by linked selection. Importantly, while the choice of \( \pi_0 \) affects \( B_d \), it does not influence the correlation between \( B_d \) and \( B \) calculated from the model \( (B_c) \), as it is only a scaling parameter. For five chromosomes, recombination rate was not known for some windows at the chromosome ends. For chromosome 8, these were 22 200-kb windows, which were excluded from all analyses. For chromosomes 1, 6, 18, and 28, this only concerned 1-3 windows, for which recombination rates were interpolated from adjacent windows. Regions with interpolated recombination rates are marked in Figure S1.

**Simulations**

To validate the models, we used individual-based simulations in fwdpp 0.4 (Thornton, 2014) with the following scenario. We assumed a constant population size of 2,000 diploid
individuals. The mutation rate was scaled accordingly to get a similar \( \pi_0 \) as the data. The probability of a recombination event at a certain position was determined by the local recombination rate, taken from Kawakami et al. (2014). Individuals for reproduction were sampled proportional to their fitness, with a constant population size (soft selection). Mating occurred by randomly pairing of gametes after mutation and recombination. As above, the density of conserved sites in a window was used to infer the probability of deleterious mutations. We simulated entire chromosomes corresponding to the sizes in the collared flycatcher reference genome and ran simulations for 6N generations. To calculate \( \pi \), we drew a random sample of 1 % of the individuals from the population.

**Data Availability Statement**

All data used for this study was previously published. See Kawakami et al. (2014), Burri et al. (2015) and Craig et al. (2017). Supplemental Material and C++ code used for the simulations and calculation of B is available at FigShare [https://figshare.com/s/5ee30bae0acfad563865](https://figshare.com/s/5ee30bae0acfad563865)

**RESULTS**

**Correlations between \( \pi \) and recombination rate and density of conserved sites**

The association between nucleotide diversity and rate of recombination in the collared flycatcher is best explained by a curvilinear relationship \( (r^2 = 0.08 \) with a power regression versus \( r^2 = 0.02 \) with a linear regression, Fig. 1a). Nucleotide diversity increased with recombination rate up to \( \approx 1.5 \) cM/Mb, but was then saturated. The variance in \( \pi \) among windows was higher at low than at high recombination rates. As expected, \( \pi \) was negatively correlated with the density of conserved sites \( (\text{dcs}; r^2 = 0.22, \) linear regression, Fig 1b). The correlation was weaker when exon density was used as a proxy for the density of target sites for selection \( (r^2 = 0.06) \), indicating an important role of selection in non-coding regions (c.f.
Craig et al., 2017). Note however, that the correlation is not exclusively due to linked selection but also due to direct selection. A multiple linear regression of $\pi$ in dependence of dcs and recombination rate (linear factors) gave $r^2=0.25$. Plots of the distribution of $\pi$, recombination rate and dcs along chromosomes readily demonstrate that the overall shape of the diversity landscape resembles the recombination landscape with diversity valleys corresponding to regions with low recombination, while fine-scale differences in $\pi$ tend to be due to variation in dcs (Fig. S1).

**Modelling of linked selection**

The genome-wide correlation between $B_c$ from model 1 (background selection) and $B_d$ in 200-kb windows was $r^2=0.28$ (Fig. 2a). Correlations for individual chromosomes are given in Table S1. An example of the relationship between $B_c$ and $B_d$ using model 1 along a chromosome (chromosome 12) is shown in Fig. 2b (for other chromosomes, see Fig. S5).

While the base line of the modeled diversity landscape closely matches the data, the amount of reduction in the diversity valley of this chromosome deviates strongly. This was the case for basically all diversity valleys, of which there usually are one or two per chromosome (see Ellegren et al. 2012 and Burri et al. 2015). Higher mutation rate or lower recombination rate than the estimates used could explain part of the difference between $B_d$ and $B_c$ in valleys (see Supplementary Material File S1).

If we include recent strong positive selection as indicated by non-synonymous substitutions fixed in the collared flycatcher lineage since its split from pied flycatcher (model 2), the overall correlation between $B_d$ and $B_c$ for $s=0.1$ was $r^2=0.45$ (Fig. 2a). As the number of sweeps modeled is likely to be higher than the actual number, model 2 underestimates $B$ more.
often and has higher variance than model 1, but the correlation is much better than with model
1. This was consistent for most chromosomes (Table S1). $B_c$ now showed deep and
pronounced diversity valleys (Fig. 2b). With lower $s_b$, the valley in chromosome 12 is less
deep and the correlation is lower but still higher than with BGS alone ($r^2=0.55$; for $s_b=0.01,$
$r^2=0.58$; for $s_b=0.05$, $r^2=0.81$; for $s_b=0.1$ $r^2=0.85$.

Model 3 incorporates sweeps at sites of substitutions since the split from the more distantly
related red-breasted flycatcher. The overall correlation using $\alpha=0.2$ for the ancestral sweeps
was weaker ($r^2=0.25$) than with both model 1 and 2. Nevertheless, as exemplified for the
terminal diversity valley of chromosome 12, some dips were only explained by model 3 (Fig.
3). For some chromosomes, model 3 in fact gave the best fit (e.g., $r^2=0.47$ in chromosome 19
vs. $r^2=0.03$ with model 2; Fig. 4, Table S1). This is important since it means that while
sweeps are needed to explain diversity valleys, they do not necessarily have to be recent.
Using $\alpha=1$ like for recent sweeps, the correlation becomes lower for chromosome 12 ($r^2=0.51$
vs 0.76 with $\alpha=0.2$).

To summarize, which of the three models of linked selection best explains the observed local
reductions in diversity differs along the genome. While BGS largely explains the base level of
genomic diversity, the $\pi$ valleys are explained better with either recent or ancient sweeps.
Across all chromosomes, 56% of the $\pi$ valleys were explained by model 2 (including recent
sweeps), 26% by model 3 (also including ancient sweeps), 9% by model 1 (only BGS), while
9% were not covered by any of the models. See Figure S5 for a comparison across all
chromosomes.
**Simulations**

Individual-based simulations of diversity along one example chromosome confirmed our analytical results. Even though the general shape of the diversity landscape and also the base level of simulated B (\(B_j\)) closely matched the analytical expectation under BGS (\(B_d\)), the diversity valley was not visible if we only allowed for deleterious mutations (Fig. 5 blue line). If in addition beneficial mutations occurred, a diversity valley emerged (Fig. 5 green line), though not as pronounced as in the empirical data. Importantly, the valley formed even though beneficial mutations occurred randomly all over the chromosome and not only in the valley region.

**DISCUSSION**

Both the simulations and the analytical results show that background selection can explain a large part of the base level and small-scale variation of genetic diversity in the flycatcher genome. The correlation between our background selection model and \(\pi\) from the data lies within the range of correlations found in similar studies in human (McVicker et al., 2009) and *Drosophila* (Comeron, 2014, Elyashiv et al., 2016). On the other hand, the deep diversity valleys present in the data cannot be explained with BGS alone. When we included selective sweeps at candidate sites in the model, or when we added 1% positively selected mutations to the simulations, we generally obtained a better fit to the empirical data. However, positive selection in the model needs to be relatively high to generate valleys as pronounced as observed in the data. There are alternative explanations to why BGS alone, with the parameters used herein, does not explain the full magnitude of variation in diversity levels. Since the rate and fitness effects of newly arising deleterious mutations are key parameters in determining the strength of BGS, a major limitation of this study might be the restriction to consider only point
mutations. Other mutation events, such as short insertions or deletions, might contribute a substantial part of the total input of deleterious mutations (e.g. Mills et al. 2006). Additionally, we assumed a distribution of fitness effects for deleterious mutations that was estimated based on non-synonymous sites and might not accurately represent the distribution of effects of other mutation types. However, as discussed in the Supplementary Material (File S1), while other reasons than selective sweeps, like under- or overestimation of used parameters (density of functional sites, mutation and recombination rate) or demography, may contribute to the depth of diversity valleys, we found none that would likely explain it exclusively. It is interesting to consider that while other forces than linked selection could increase the valley depth, they also would likely distort the correlation of \( \pi \) with the linked selection landscape formed by recombination and density of functional sites. Additionally, there are several modes of selection, like balancing selection, that our models did not consider. As we only used non-synonymous substitutions as candidate sites for positive selection, it is also possible that additional sweeps affected non-coding regions, which might result in diversity valleys not explained by any of our models. These could be possible explanations for the valleys that could not be explained by our models. Our results concerning the relative contributions of BGS and hitchhiking to linked selection go in the same direction with the findings of Elyashiv et al. (2016) in *Drosophila melanogaster*. For 100-kb windows, they found a correlation between \( \pi \) and \( B \) with BGS + sweeps of 0.44, compared to 0.42 with only BGS. The difference between the models with and without sweeps was thus larger in flycatchers than *Drosophila*. Notably, the valleys in the heterogeneous diversity landscape so characteristic for flycatchers was better explained with sweeps at sites of non-synonymous substitutions in the flycatcher lineage than with BGS alone. The simulations showed that even with a random location of sweeps (i.e., not restricted to sites of non-synonymous substitutions), diversity valleys were better captured than with BGS alone (see Fig. 5). This can be explained by the higher impact of positive selection
in regions with low recombination. With a higher overall effect of linked selection, recombination rate has a higher impact on the diversity landscape, leading to the formation of pronounced diversity valleys in low recombination regions.

In general, model 3 (with recent and ancient sweeps) gave a worse fit than model 2 (recent sweeps). This is because more false sweeps are introduced, producing valleys that do not fit the data and negatively affect the correlation. Although the model with recent sweeps was better than other models in explaining diversity valleys, in some cases only the model that also incorporated ancient sweeps could explain the presence of such valleys (see Fig. 3 and 4). That sweeps do not need to be recent or specific to the focal species for generating locally reduced diversity levels is consistent with the finding that the locations of valleys are often conserved among *Ficedula* species (Burri et al. 2015). Most diversity valleys are located at chromosome ends and or at the position of presumed centromeres (Ellegren et al. 2012; Burri et al., 2015). Sweeps could occur relatively frequently in these regions due to a meiotically driven arms race in centromeres or telomeres (Henikoff et al., 2001, Malik & Henikoff 2009). A role, yet not exclusive, of positive selection behind the heterogeneous flycatcher diversity landscape is also consistent with the observation that Fay & Wu‘s H statistic often shows signatures of positive selection in diversity valleys (Burri et al., 2015).

Contrary to the expectations for speciation islands, our findings indicate that diversity valleys in flycatchers are not a direct consequence of locally reduced gene flow. However, the reduced effective population size caused by the effects of linked selection in these regions might still promote the rapid build-up of drift-induced Dobzhansky-Muller incompatibilities or other isolating mechanisms during an allopatric phase, therefore reducing subsequent gene flow in secondary contact (Dobzhansky, 1936; Muller, 1942). Thus, although reduced gene flow can
be safely rejected as cause for differentiation peaks in flycatchers, fully rejecting them as 'speciation islands’ might be premature.

Our results indicate that positive selection contributes to the base variation of genetic diversity across the genome rather than producing classical outliers. Recombination rate and the (spatial and temporal) abundance of selected mutations determine how selection influences nearby neutral sites. This is true for deleterious and beneficial mutations, though the impact varies. It is thus quite reasonable to conclude that recombination rate and functional site density form a linked-selection scaffold which determines the impact of linked selection in a certain region of the genome. This is the shape we can calculate with equation 1, while the total impact depends on selection strength and sign. This may be one explanation why selective sweeps are so hard to find: they are embedded in the general linked-selection landscape. But it is important to keep in mind that also the reverse is true: if there are diversity valleys in the data, it does not necessarily mean that there is something special happening - the impact of linked selection might just be stronger in this region. This is also true for summary statistics which depend on linked selection. Note that even for windows without sites under selection we do not find $B=1$, which means that the whole genome is influenced by linked selection. Thus, we should be careful when using intergenic regions as neutral baseline. Our results thus add to the general demand for using background selection as a null hypothesis for diversity levels (Comeron 2017).

ACKNOWLEDGEMENTS

This research was supported by the Swedish Research Council (2013-8271) and Knut and Alice Wallenberg foundation. AN was financially supported by a Swiss National Science Foundation
grant (P300PA_177852). We thank Reto Burri and Rory Craig for sharing their data files and David Castellano and two anonymous reviewers for helpful comments.

REFERENCES

Baum, L.E.; Petrie, T.; Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Annals of Mathematical Statistics 41(1), 164–171. doi: 10.1214/aoms/1177697196

Begun, D. J. and Aquadro, C. F. (1992). Levels of naturally occurring DNA polymorphism correlate with recombination rates in d. melanogaster. Nature, 356(6369):519–520.

Bolivar, P., Mugal, C. F., Rossi, M., Nater, A., Wang, M., Dutoit, L., & Ellegren, H. (2018). Biased inference of selection due to GC-biased gene conversion and the rate of protein evolution in flycatchers when accounting for it. Molecular biology and evolution, 35(10), 2475-2486.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bureš, S., Garamszegi, L. Z., 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(11):1656–1665.

Charlesworth, B. (1996). Background selection and patterns of genetic diversity in Drosophila melanogaster. Genetics Research, 68(2):131–149.

Charlesworth, B. (1998). Measures of divergence between populations and the effect of forces that reduce variability. Molecular biology and evolution, 15(5), 538-543.

Charlesworth, B., Morgan, M., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics, 134(4):1289–1303.

Comeron, J. M. (2014). Background selection as baseline for nucleotide variation across the drosophila genome. PLoS Genetics, 10(6): e1004434.

Comeron, J. M. (2017). Background selection as null hypothesis in population genomics: insights and challenges from drosophila studies. Philosophical Transactions of the Royal Society of London B, 372(1736):20160471.
Craig, R., Suh, A., Wang, M. & Ellegren, H. (2018) Natural selection beyond genes: Identification and analyses of evolutionarily conserved elements in the genome of the collared flycatcher (*Ficedula albicollis*). *Molecular Ecology* 27:476-492

Crow, J. F., & Kimura, M. (1970). *An introduction to population genetics theory*. New York, Harper & Row

Cruickshank, T. E. and Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology, 23*(13):3133–3157.

Cutter, A. D. and Payseur, B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nature Reviews Genetics, 14*(4):262–274.

Dobzhansky, T. H. (1936). Studies on hybrid sterility. II. Localization of sterility factors in Drosophila pseudoobscura hybrids. *Genetics, 21*(2), 113.

Dutoit, L., Vijay, N., Mugal, C. F., Bossu, C. M., Burri, R., Wolf, J., and Ellegren, H. (2017). Covariation in levels of nucleotide diversity in homologous regions of the avian genome long after completion of lineage sorting. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 284*(1849), 20162756.

Elyashiv, E., Sattath, S., Hu, T. T., Strustovsky, A., McVicker, G., Andolfatto, P., Coop, G., and Sella, G. (2016). A genomic map of the effects of linked selection in drosophila. *PLoS Genetics, 12*(8), e1006130.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nature Reviews Genetics, 8*(8):610–618.

Fay, J. C. and Wu, C.-I. (2000). Hitchhiking under positive darwinian selection. *Genetics, 155*(3):1405–1413.

Feder, J. L. and Nosil, P. (2010). The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. *Evolution, 64*(6):1729–1747.

Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. *Science, 293*(5532):1098–1102.

Hudson, R. R. and Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics, 141*(4):1605–1617.

Kaplan, N. L., Hudson, R., and Langley, C. (1989). The" hitchhiking effect" revisited. *Genetics, 123*(4):887–899.

Kawakami, T., Smeds, L., Backström, N., Husby, A., Qvarnström, A., Mugal, C. F., Olason, P., and 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. 

*Molecular Ecology*, 23(16):4035–4058.

Kawakami T, Mugal CF, Suh A, Nater A, Burri R, Smeds L, Ellegren H. (2017) Whole-genome patterns of linkage disequilibrium across flycatcher populations clarify the causes and consequences of fine-scale recombination rate variation in birds. Molecular Ecology 26:4158–4172.

Kim, Y. and Stephan, W. (2000). Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics*, 155(3):1415–1427.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777.

Korneliussen, T.S., Albrechtsen, A., and Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15, 356

Lercher, M. J. and Hurst, L. D. (2002). Human snp variability and mutation rate are higher in regions of high recombination. *Trends in Genetics*, 18(7):337–340.

Lundberg, A. and Alatalo, R. (1992). The pied flycatcher london. *UK: Poyser*.

Mackay, T. F., Richards, S., Stone, E. A., Barbadilla, A., Ayroles, J. F., Zhu, D., ... & Richardson, M. F. (2012). The Drosophila melanogaster genetic reference panel. *Nature*, 482(7384), 173.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, 23(1):23–35.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genetics*, 5(5):e1000471.

Mills RE, Luttig CT, Larkins CE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*. 2006;16(9):1182-90.

Muller, H. (1942). Isolating mechanisms, evolution, and temperature. In *Biol. Symp. (Vol. 6, pp. 71-125)*.

Nachman, M. W. (2001). Single nucleotide polymorphisms and recombination rate in humans. *Trends in Genetics*, 17(9):481–485.

Nater, A., Burri, R., Kawakami, T., Smeds, L., and Ellegren, H. (2015). Resolving evolutionary relationships in closely related species with whole-genome sequencing data. *Systematic Biology*, 64(6), 1000-1017.

Noor, M. A. and Bennett, S. M. (2009). Islands of speciation or mirages in the desert? examining the role of restricted recombination in maintaining species. *Heredity*, 103(6):439–444.
Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research, 67*(2):159–174.

Nordborg, M., Hu, T. T., Ishino, Y., Jhaveri, J., Toomajian, C., Zheng, H., Bakker, E., Calabrese, P., Gladstone, J., Goyal, R., et al. (2005). The pattern of polymorphism in *Arabidopsis thaliana*. *PLoS Biology, 3*(7):e196.

Nosil, P. (2008). Speciation with gene flow could be common. *Molecular Ecology, 17*:2103–2106.

Nosil, P., Egan, S. P., and Funk, D. J. (2008). Heterogeneous genomic differentiation between walking-stick ecotypes: “isolation by adaptation” and multiple roles for divergent selection. *Evolution: International Journal of Organic Evolution, 62*(2), 316-336.

Ravinet, M., Faria, R., Butlin, R., Galindo, J., Bierne, N., Rafajlović, M., Noor, M., Mehlig, B., and Westram, A. (2017). Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *Journal of Evolutionary Biology, 30*(8):1450–1477.

Samuk, K., Owens, G. L., Delmore, K. E., Miller, S., Rennison, D. J., and Schluter, D. (2017). Gene flow and selection interact to promote adaptive divergence in regions of low recombination. *Molecular ecology, 26*(17), 4378-4390.

Seehausen, O., Butlin, R. K., Keller, I., Wagner, C. E., Boughman, J. W., Hohenlohe, P. A., Peichel, C. L., Saetre, G.-P., Bank, C., Brännström, Å., et al. (2014). Genomics and the origin of species. *Nature Reviews Genetics, 15*(3):176–192.

Smeds, L., Qvarnstrom, A., and Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome research, gr*-204669.

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*(1544), 1245-1253.

Tenaillon, M. I., Sawkins, M. C., Long, A. D., Gaut, R. L., Doebley, J. F., and Gaut, B. S. (2001). Patterns of dna sequence polymorphism along chromosome 1 of maize (zea mays ssp. mays l.). *Proceedings of the National Academy of Sciences, 98*(16):9161–9166.

Thornton, K. R. (2014). A c++ template library for efficient forward-time population genetic simulation of large populations. *Genetics, 198*:157–166.

Turner, T. L. and Hahn, M. W. (2010). Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology, 19*(5):848–850.
Van Doren, B. M., Campagna, L., Helm, B., Illera, J. C., Lovette, I. J., & Liedvogel, M. (2017). Correlated patterns of genetic diversity and differentiation across an avian family. *Molecular ecology, 26*(15), 3982-3997.

Wolf, J. B. and Ellegren, H. (2017). Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics, 18*(2), 87.
Figure 1: Relationship between mean nucleotide diversity and (a) recombination rate in $cM/Mb$ ($r^2=0.08$) and (b) proportion of conserved sites ($r^2=0.22$) for 200-kb windows.
Figure 2: a) Genome-wide correlation between $B_c$ (the modeled deviation from neutral diversity) and $B_d$ ($B$ calculated from data) under two different linked selection models. The gray line marks identity for better orientation. Colored lines are linear regressions with blue showing only background selection (model 1) and green showing BGS and recent sweeps (model 2). b) A comparison between $B_d$ (black) and $B_c$ calculated with BGS (model 1, blue line) and BGS plus recent sweeps (model 2, green), with $s_b=0.1$, $\alpha=1$ for chromosome 12.
Figure 3: A comparison between $B_d$ (black) and $B_c$ calculated with BGS plus recent ($s_b=0.1$, $\alpha=1$) and ancient ($s_b=0.1$, $\alpha=0.2$) sweeps (model 3) (purple). The grey line represents $B_c$ under model 1. Parameters for recent sweeps as in Fig. 2.

Figure 4: A comparison between $B_d$ (black) and $B_c$ from model 1 (blue), model 2 (green), and model 3 (purple) for chromosome 19. Parameters as Fig. 2.
Figure 5: Simulation of chromosome 12, population size $N=2000$, run for $6N$ generations, selection coefficient for deleterious mutations $s_d=-0.001$, averaged over 40 runs. Blue line: only background selection; green line: BGS plus 1% beneficial mutations with $s_b=0.01$. The black line shows $B$ calculated from the empirical data.

Table 1: Model parameters

| Parameter | Value | Reference                |
|-----------|-------|--------------------------|
| $\pi_0$   | 0.0048| This study               |
| $\mu$     | $4.6 \times 10^{-9}$ | Smeds et al. (2016)     |
| $r$       | per window | Kawakami et al. (2014) |
| $N_e$     | 450,000 | Nater et al. (2015)      |
| $T$       | 320,000 | Nater et al. (2015)      |
| $u_d$     | per window | Depends on dcs, Craig et al. (2017) |
| $s_d$     | DFE   | Bolivar et al. 2018      |
| $s_b$     | 0.1/0.05/0.01 | Variable model parameter |
| $\alpha$  | 1/0.2 | Variable model parameter |