Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers

Reto Burri, Alexander Nater, Takeshi Kawakami, et al.

Genome Res. published online September 9, 2015
Access the most recent version at doi:10.1101/gr.196485.115

Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers

Reto Burri[1*], Alexander Nater[1], Takeshi Kawakami[1], Carina F. Mugal[1], Pall I. Olason[2], Linnea Smeds[1], Alexander Suh[1], Ludovic Dutoit[1], Stanislav Bureš[3], Laszlo Z. Garamszegi[4], Silje Hogner[5,6], Juan Moreno[7], Anna Qvarnström[8], Milan Ružić[9], Stein-Are Sæther[5,10], Glenn-Peter Sætre[5], Janos Török[11] & Hans Ellegren[1*]

[1] Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, 75236 Uppsala, Sweden; [2] Wallenberg Advanced Bioinformatics Infrastructure (WABI), Science for Life Lab, Uppsala University, 75123 Uppsala, Sweden; [3] Laboratory of Ornithology, Department of Zoology, Palacky University, 77146 Olomouc, Czech Republic; [4] Department of Evolutionary Ecology, Estación Biológica de Doñana-CSIC, 41092, Seville, Spain; [5] Department of Biosciences, Centre for Ecological and Evolutionary Synthesis, University of Oslo, 0316 Oslo, Norway; [6] Natural History Museum, University of Oslo, 0318 Oslo, Norway; [7] Museo Nacional de Ciencias Naturales-CSIC, 28006 Madrid, Spain; [8] Department of Animal Ecology, Evolutionary Biology Centre, Uppsala University, 75236 Uppsala, Sweden; [9] Bird Protection and Study Society of Serbia, Radnička 20a, 21000 Novi Sad, Serbia; [10] Norwegian Institute for Nature Research (NINA), 7034 Trondheim, Norway; [11] Behavioural Ecology Group, Department of Systematic Zoology and Ecology, Eötvös Loránd University, 1117 Budapest, Hungary.

[*] Correspondence to:
reto.burri@ebc.uu.se, Tel +46 70 534 00 71, Fax +46 18 471 63 10 (RB)
hans.ellegren@ebc.uu.se, Tel +46 18 471 64 60, Fax +46 18 471 63 10 (HE)

Running Title: Evolution of Genome Differentiation

Keywords: background selection, genomic islands, selective sweeps, speciation

Abstract

Speciation is a continuous process during which genetic changes gradually accumulate in the genomes of diverging species. Recent studies have documented highly heterogeneous differentiation landscapes, with distinct regions of elevated differentiation ('differentiation islands') widespread across genomes. However, it remains unclear which processes drive the evolution of differentiation islands, how the differentiation landscape evolves as speciation advances, and ultimately how differentiation islands are related to speciation. Here, we addressed these questions based on population genetic analyses of 200 re-sequenced genomes from 10 populations of four *Ficedula* flycatcher sister species. We show that a heterogeneous differentiation landscape starts emerging among populations within species and that differentiation islands evolve recurrently in the very same genomic regions among independent lineages. Contrary to expectations from models that interpret differentiation islands as genomic regions involved in reproductive isolation that are shielded from gene flow, patterns of sequence divergence (d_{xy} and relative node depth) do not support a major role of gene flow in the evolution of the differentiation landscape in these species. Instead, as predicted by models of linked selection, genome-wide variation in diversity and differentiation can be explained by variation in recombination rate and the density of targets for selection. We thus conclude that the heterogeneous landscape of differentiation in *Ficedula* flycatchers evolves mainly as the result of background selection and selective sweeps in genomic regions of low recombination. Our results emphasize the necessity of incorporating linked selection as a null model to identify genome regions involved in adaptation and speciation.

Introduction

Uncovering the genetic architecture of reproductive isolation and its evolutionary history are central tasks in evolutionary biology. The identification of genome regions that are highly differentiated between closely related species, and thereby constitute candidate regions involved in reproductive isolation, has recently been a major focus of speciation genetic research. Studies from a broad taxonomic range, involving organisms as diverse as plants (Green et al. 2010; Renaut et al. 2013), insects (Turner et al. 2005; Lawniczak et al. 2010; Nadeau et al. 2012; Soria-Carrasco et al. 2014), fishes (Jones et al. 2012), mammals (Harr 2006), and birds (Ellegren et al. 2012) contribute to the emerging picture of a genomic landscape of differentiation that is usually highly heterogeneous, with regions of locally elevated differentiation ('differentiation islands') widely spread over the genome. However, the evolutionary processes driving the evolution of the differentiation landscape and the role of differentiation islands in speciation are subject to controversy (Turner and Hahn 2010; Cruickshank and Hahn 2014; Pennisi 2014).

Differentiation islands were originally interpreted as 'speciation islands', regions that harbour genetic variants involved in reproductive isolation and are shielded from gene flow by selection (Turner et al. 2005; Soria-Carrasco et al. 2014). During speciation-with-gene-flow, speciation islands were suggested to evolve through selective sweeps of locally adapted variants and by hitchhiking of physically linked neutral variation ('divergence hitchhiking', Via and West 2008); gene flow would keep differentiation in the remainder of the genome at bay (Nosil 2008; Nosil et al. 2008). In a similar way speciation islands can arise by allopatric speciation followed by secondary contact. In this case, genome-wide differentiation increases during periods of geographic isolation, but upon secondary contact it is reduced by gene flow in genome regions not involved in reproductive isolation. In the absence of gene flow in allopatry, speciation islands need not (but can) evolve by local adaptation, but may consist of intrinsic incompatibilities *sensu* Bateson-Dobzhansky-Muller(Bateson 1909; Dobzhansky 1937; Muller 1940) that accumulated in spatially isolated populations.

However, whether differentiation islands represent speciation islands has been questioned. Rather than being a cause of speciation, differentiation islands might evolve only after the onset of reproductive isolation as a consequence of locally accelerated lineage sorting (Noor and Bennett 2009; Turner and Hahn 2010; White et al. 2010; Cruickshank and Hahn 2014; Renaut et al. 2014), such as in regions of low recombination (Nachman 2002; Sella et al. 2009; Cutter and Payseur 2013). In these regions the diversity-reducing effects of both positive selection and purifying selection (background selection, BGS) at linked sites ('linked selection') impact physically larger regions due to the stronger linkage among sites. The thereby locally reduced effective population size (N_e) will enhance genetic drift and hence inevitably lead to increased differentiation among populations and species.

These alternative models for the evolution of a heterogeneous genomic landscape of differentiation are not mutually exclusive, and their population genetic footprints can be difficult to discern. In the cases of (primary) speciation-with-gene-flow and gene flow at secondary contact, shared variation outside differentiation islands partly stems from gene flow. In contrast, under linked selection ancestral variation is reduced and differentiation elevated in regions of low recombination, while the remainder of the genome may still share considerable amounts of ancestral genetic variation and show limited differentiation. Many commonly used population genetic statistics do not capture these different origins of shared genetic variation and have the same qualitative expectations under both models, such as reduced diversity (π) and skews towards an excess of rare variants (e.g. lower Tajima's D) in differentiation islands relative to the remainder of the genome. However, since speciation islands should evolve by the prevention or breakdown of differentiation by gene flow in regions not involved in reproductive isolation, substantial gene flow should be detectable in these regions (Cruickshank and Hahn 2014) and manifested in the form of reduced sequence divergence (d_{xy}) or as an excess of shared derived alleles in cases of assymetrical gene flow (Patterson et al. 2012). Under linked selection predictions are opposite for d_{xy} (Cruickshank and Hahn 2014), owing to reduced ancestral diversity in low-recombination regions. Further predictions for linked selection include positive

and negative relationships of recombination rate with genetic diversity (π) and differentiation (F_{ST}), respectively, and inverse correlations of the latter two with the density of targets for selection. Finally, important insights into the nature of differentiation islands may be gained by studying the evolution of differentiation landscapes across the speciation continuum.

Theoretical models and simulations of speciation-with-gene-flow predict that after an initial phase during which differentiation establishes in regions involved in adaptation, differentiation should start spreading from these regions across the entire genome (Feder et al. 2012; Flaxman et al. 2013; Feder et al. 2014).

Unravelling the processes driving the evolution of the genomic landscape of differentiation, and hence understanding how genome differentiation unfolds as speciation advances, requires genome-wide data at multiple stages of the speciation continuum and in a range of geographical settings from allopatry to sympatry (Seehausen et al. 2014). Although studies of the speciation continuum are emerging (Hendry et al. 2009; Kronforst et al. 2013; Shaw and Mullen 2014 and references therein), empirical examples of genome differentiation at multiple levels of species divergence remain scarce (Andrew and Rieseberg 2013; Kronforst et al. 2013; Martin et al. 2013), and, to our knowledge, have so far not jointly addressed the predictions of alternative models for the evolution of the genomic landscape of differentiation.

In the present study, we implemented such a study design encompassing multiple populations of four black-and-white flycatcher sister species of the genus *Ficedula* (**Fig. 1a, 1b, Supplemental Fig. S1**; for a comperehensive reconstruction of the species tree see Nater et al. in press). Previous analyses in collared flycatcher (*F. albicollis*) and pied flycatcher (*F. hypoleuca*) revealed a highly heterogeneous differentiation landscape across the genome (Ellegren et al. 2012). An involvement of gene flow in its evolution would be plausible, as hybrids between these species occur at low frequencies in sympatric populations in eastern Central Europe and on the Baltic Islands of Gotland and Öland (Alatalo et al. 1990; Sætre et al. 1999), although a recent study based on genome-wide markers identified no hybrids beyond the F_1 generation (Kawakami et al. 2014a). Still, gene flow from pied into collared flycatcher appears to have occurred (Borge et al.

2005; Backström et al. 2013; Nadachowska-Brzyska et al. 2013) despite premating isolation (reviewed in Sætre and Sæther 2010), hybrid female sterility (Alatalo et al. 1990; Tegelström and Gelter 1990), and strongly reduced long-term fitness of hybrid males (Wiley et al. 2009). Atlas flycatcher (*F. speculigera*) and semicollared flycatcher (*F. semitorquata*) are two closely related species, which have been less studied, but may provide interesting insights into how genome differentiation evolves over time. Here, we take advantage of this system to identify the processes underlying the evolution of differentiation islands based on the population genetic analysis of whole-genome re-sequencing data of 200 flycatchers.

Results

Recurrent evolution of a highly similar genomic landscape of differentiation in independent lineages

We re-sequenced genomes from four populations each of collared flycatcher and pied flycatcher, one population each of Atlas flycatcher and semicollared flycatcher (n=20 of all populations), and single individuals of two outgroup species, red-breasted flycatcher (*F. parva*) and snowy-browed flycatcher (*F. hyperythra*) (**Supplemental Table S1**). The average sequencing coverage was 14.7 ± 4.5 x (mean ± SD) (**Supplemental Table S2**) and we identified a total of about 50 million variant sites across the four focal species (73 million variant sites including outgroups). On average, two random chromosomes from any two species differ by 0.448 % to 0.502 %, which is only 1.13 to 1.68 times the average number of differences observed between two chromosomes within species (average nucleotide diversity, π: collared, $3.95 \cdot 10^{-3}$; pied $3.20 \cdot 10^{-3}$; Atlas, $3.06 \cdot 10^{-3}$; semicollared, $2.98 \cdot 10^{-3}$). As illustrated by the distributions of genome-wide differentiation (F_{ST}) among populations and species, our data cover a wide range of differentiation across the speciation continuum (**Fig. 1d**).

In line with results from a pilot comparison of ten collared and ten pied flycatchers that identified about 50 differentiation islands distributed across most chromosomes (Ellegren et al.

2012), all pair-wise comparisons between species revealed highly heterogeneous genomic landscapes of differentiation (**Fig. 1c, 1e, Supplemental Fig. S2, S3**). Strikingly, differentiation along the genomes was highly correlated between all possible pair-wise comparisons among species (**Fig. 1c, 1e, Supplemental Fig. S2, S3**). Since these comparisons sample some branches repeatedly and therefore are not phylogenetically independent, we estimated population branch statistics (*PBS*), which represent lineage-specific F_{ST} (Shriver et al. 2004; Yi et al. 2010). This analysis revealed highly concordant differentiation landscapes among species, with islands independently appearing in the same genomic regions across all species (**Fig. 1c, 1e, Supplemental Fig. S2, S3**). The first axis (PC1) of a principal component analysis (PCA) captured 97.8% of the variance in *PBS* of homologous 200 kb windows among lineages. Moreover, F_{ST} in two phylogenetically independent pairwise comparisons, pied *vs.* Atlas flycatcher, and collared *vs.* semicollared flycatcher (**Fig. 1a**) were highly correlated (R=0.804, t=96.6, $p<10^{-15}$; PC1 explaining 99.6% of the common variance; **Fig. 2a**). Furthermore, differentiation from outgroup species was also increased in the regions corresponding to differentiation islands among the black-and-white flycatchers (**Fig. 1e**, **Supplemental Fig. S3**). The findings were independent of window size, with qualitatively congruent results with a window size of 5 kb.

These results demonstrate recurrent evolution of high differentiation in the same genomic regions across the speciation continuum and may be best explained by shared genomic features underlying the emergence of differentiation islands in independent lineages. In contrast, a parallel involvement of the very same large set of loci in reproductive isolation in different lineages appears unlikely. The former is supported by reduced levels of nucleotide diversity (π) in differentiation islands of all species (**Supplemental Table S3**) and a high correlation of π among species (**Fig. 1c, 1e**, **Supplemental Fig. S2**; PC1 explains 93.6% of the genome-wide variation), suggesting common underlying mechanisms reducing N_e in differentiation islands. Importantly, a reduction of N_e in a common ancestor is not sufficient to explain the pattern of locally reduced diversity seen in all species, as this would be incompatible with high F_{ST} in those

regions in comparisons between descendent lineages. Still, to further explore this possibility we focused on SNPs unique to each species, which are expected to be highly enriched for lineage-specific mutations affected only by processes in descendent lineages. We found that lineage-specific SNPs showed strongly reduced diversity in differentiation islands in all four species (**Fig. 1c**, **Supplemental Fig. S2**, **Supplemental Table S3**), evidencing recurrent processes contributing towards highly concordant differentiation landscapes among species. Moreover, differentiation islands were not only observed in comparisons between species. Similar to *Heliconius* butterflies (Martin et al. 2013), already in the geographically most isolated populations within collared and pied flycatchers, regions corresponding to differentiation islands between species exhibited higher F_{ST} within species than the genomic background of low differentiation (**Fig. 1c, 1e**, **Supplemental Fig. S4**). About half of the genomic windows within interspecific differentiation islands showed elevated differentiation also in any of the comparisons among collared (46.8%) or pied flycatcher (49.6%) populations, and within-species differentiation of the geographically most isolated populations was positively correlated with interspecific differentiation (*PBS* collared Italy: t=20.6, $p<10^{-15}$; R^2=0.084; *PBS* pied Spain: t=30.5, $p<10^{-15}$; R^2=0.167).

Gene flow is not a main factor governing the heterogeneous differentiation landscape

To investigate the predicted characteristics of the speciation island model, we explored whether the gene flow previously reported between collared and pied flycatcher (Backström et al. 2013; Nadachowska-Brzyska et al. 2013) may have homogenized genomes after secondary contact following the last glaciation. In such a scenario differentiation is expected to decrease with increasing opportunity for gene flow (Martin et al. 2013) and an excess of derived variants shared between sympatric populations should be observed. To test these predictions, we benefitted from a study design that included four pairs of collared flycatcher and pied flycatcher populations (**Figure 1b, Supplemental Table S1**) that differ in the degree and timing of

interspecific contact: (i) the presumed refugial populations of the respective species in Italy and Spain, (ii) allopatric populations from Hungary and Sweden, (iii) populations from an old zone of sympatry in the Czech Republic, and (iv) populations from recent sympatry (50-150 years ago, Lundberg and Alatalo 1992) on the Baltic Island of Öland. Analysis of differentiation patterns showed that, against the prediction of a model with postglacial gene flow, sympatric populations were not less differentiated than allopatric populations (mean F_{ST}, sympatric: Czech Republic, 0.291, Baltic Sea island, 0.303; allopatric: Italy-Spain, 0.274; Hungary-Sweden, 0.288). This result was further supported by ABBA-BABA tests that found no excess of shared derived variants among sympatric populations of the two species (Tests 1-2, **Supplemental Table S4**).

To address the possibility of gene flow during previous interglacial periods we first used ABBA-BABA tests with collared and pied flycatcher represented by their presumable refugial populations from Italy and Spain, respectively, together with the other two black-and-white flycatchers. Footprints of gene flow during this period were detected among collared, pied, and semicollared flycatchers (Tests 3-6, **Supplemental Table S4**), potentially indicating an impact of ancient gene flow on genome differentiation. However, this was not supported by an analysis of the variation in sequence divergence (d_{xy}) along the genome. F_{ST} and d_{xy} should both be reduced if variation in the genomic background is homogenized by gene flow, while differentiation islands are expected to show higher F_{ST} and d_{xy} if they are resistant to introgression (Cruickshank and Hahn 2014). The same predictions as for d_{xy} apply to relative node depth (RND, Feder et al. 2005), which aims at correcting sequence divergence for mutation rate variation. Such a pattern was not observed in flycatchers (**Fig. 1c**, **Supplemental Fig. S2, Supplemental Tables S5, S6**). In contrary, differentiation islands exhibited lower d_{xy} and *RND* than the genomic background, suggesting reductions of N_e in islands already in the ancestral population (Nachman and Payseur 2012) that exceed the amplitude of potential antagonistic effects of gene flow on sequence divergence. Finally, contrary to predictions from models of genetic hitchhiking during speciation-with-gene-flow (Feder et al. 2012; Flaxman et al. 2013), no sign of broadening of differentiation islands across the flycatcher speciation continuum was

observed (**Fig. 1c**, **Supplemental Fig. S2**), not even in comparisons including the two outgroup species (**Fig. 1e**, **Supplemental Fig. S3**).

Accelerated lineage sorting in regions of low recombination

An alternative hypothesis to the idea that differentiation islands are speciation islands is that they represent genomic regions of low intraspecific rates of recombination in which the effect of linked selection on N_e, and therefore levels of diversity, is pronounced (Noor and Bennett 2009; Nachman and Payseur 2012). The access to a recently developed high-density linkage map of the collared flycatcher genome (Kawakami et al. 2014a; Kawakami et al. 2014b) enabled us to examine the relationship between recombination and differentiation in more detail than previously possible. Lineage-specific differentiation (*PBS*) significantly increased with decreasing recombination rate (**Fig. 2b**; linear regression, t=-40.7, $p<10^{-15}$, $R^2=0.266$) and low recombination was a hallmark of differentiation islands (marked examples including, for example, Chromosomes 4A, 10, 12, and 17, **Fig. 1c**, **Supplemental Fig. S2**). Importantly, differentiation evolved faster towards fixation of segregating variants in low-recombination regions across the speciation continuum (**Fig. 1d**).

Strong impact of linked selection on the evolution of the differentiation landscape

As predicted for a model of linked selection (Charlesworth et al. 1993; Sella et al. 2009), genetic diversity (π) significantly correlated with recombination rate and the density of targets for selection measured as the density of coding sequence (multiple linear regression, recombination rate: t=2.923, p=0.003; exon density: t=-17.3, $p<10^{-15}$; interaction: t=7.5, $p<10^{-12}$; $R^2=0.131$). The same result was found for d_{xy} (multiple linear regression, recombination rate: t=-3.5, $p=5\cdot10^{-4}$; exon density: t=-19.9, $p<10^{-15}$; interaction: t=9.8, $p<10^{-15}$; $R^2=0.121$; without interaction: recombination rate: t=8.1, $p<10^{-15}$; exon density: t=-22.4, $p<10^{-15}$; $R^2=0.102$) and *RND* (multiple

linear regression, recombination rate: t=4.1, $p=4.5 \cdot 10^{-5}$; exon density: t=-11.1, $p<10^{-15}$; interaction: t=9.1, $p<10^{-15}$; $R^2=0.102$), which at this level of divergence mostly reflect ancestral diversity. As predicted by the dependence of differentiation on within-population diversity (Charlesworth 1998), differentiation was highly negatively correlated to diversity (linear regression, t=-71.3, $p<10^{-15}$, $R^2=0.523$). Accordingly, differentiation correlated with recombination rate and the density of targets for selection (multiple linear regression, recombination: t=-21.3, $p<10^{-15}$; exon density: t=6.8, $p<10^{-10}$; interaction: t=-2.7, p=0.007; $R^2=0.279$).

Tajima's *D* was reduced in differentiation islands (**Supplemental Fig. S2**, **Supplemental Table S7**), demonstrating a skew in the site frequency spectrum towards an excess of low-frequency variants. While this signal is in line with the action of linked selection in the evolution of differentiation islands, it does not discern whether linked selection occurred in the form of background selection against slightly deleterious mutations alone, or whether positive selection for advantageous variants was also involved. Moreover, selective sweeps may be difficult to detect in the presence of background selection (e.g. Enard et al. 2014). We therefore investigated the unfolded site frequency spectrum for skews that are unique to hitchhiking under positive selection. Low values of Fay and Wu's *H* (Fay and Wu 2000), indicative of an excess of high-frequency derived variants, were almost exclusively found in highly differentiated genome regions (**Fig. 1c, Supplemental Fig. S2**). However, contrary to all other population genetic parameters, drops in *H* in one species were usually not paralleled in other species (**Fig. 1c, Supplemental Fig. S2**), suggesting lineage-specific episodes of positive selection. Around five percent of conspicuous differentiation islands per species showed a pronounced excess of high-frequency derived variants expected from selective sweeps (**Fig. 3**, **Supplemental Fig. S2**, **Supplemental Table S8**). This leads us to conclude that in addition to a dominating role of background selection, divergent selection contributes to the evolution of reduced diversity and increased differentiation in a fraction of genome regions with low recombination.

Discussion

Genome-wide variation in gene flow and the effects of linked selection represent two mutually non-exclusive processes that can contribute to the evolution of heterogenous genomic landscapes of differentiation. If regions involved in reproductive isolation coincide with regions of low recombination, both processes may contribute to the evolution of low diversity and high differentiation in such regions relative to the remainder of the genome. Their effects may then only be discerned by their antagonistic effects on sequence divergence (d_{xy}), which is reduced in low recombination regions by linked selection, but reduced elsewhere by gene flow. Here, the demonstration of strongly reduced sequence divergence in low recombination regions is not compatible with a pure model of heterogeneous gene flow, but provides compelling evidence for a significant effect of linked selection on the genomic landscapes of diversity and differentiation. Therefore, even though a contribution of gene flow to the observed patterns cannot be excluded, the diversity-reducing effects of linked selection must have been the dominant driver of high differentiation in genome regions of low recombination in *Ficedula* flycatchers.

We conclude that the heterogeneous genomic landscape of differentiation in *Ficedula* flycatchers evolves mainly as a consequence of a heterogeneous landscape of recombination. It starts emerging in structured populations and, owing to the conservation of the recombination landscape among species, evolves recurrently in independent lineages across the speciation continuum due to effects of linked selection. This supports models according to which the genomic landscape of differentiation does not primarily reflect processes associated with reproduction isolation and speciation (Turner and Hahn 2010; White et al. 2010; Roesti et al. 2012; Renaut et al. 2013). However, contrary to previous assertions (Cruickshank and Hahn 2014), speciation is not required for the evolution of a heterogeneous differentiation landscape, but more moderate reductions in gene flow such as in structured populations are sufficient to trigger this process. These results imply that differentiation islands are not necessarily involved in speciation, but do not exclude the possibility that a subset of these regions harbour speciation

317 genes, or that enhanced rates of genetic drift due to locally reduced N_e facilitate the (neutral)
318 fixation of intrinsic incompatibilities in a subset of these regions.

319 The generality of the role of recombination in mediating the effects of linked selection
320 (Charlesworth and Campos 2014) suggests that our conclusions are of broad relevance to the
321 fields of speciation and adaptation genomics. Associations of high differentiation with low
322 recombination rates in species with well-documented interspecific gene flow (Andrew and
323 Rieseberg 2013; Renaut et al. 2013; Renaut et al. 2014) demonstrate that recombination rate
324 variation is an important factor in determining the distribution of differentiation across the
325 genome even in cases of speciation-with-gene-flow. When gene flow occurs, its relative impact
326 on the evolution of the genomic landscape of differentiation compared to linked selection will
327 depend on various factors, including the amplitude of recombination rate variation across the
328 genome. Everything else equal, linked selection will lead to more heterogeneous differentiation
329 landscapes in species with high recombination rate variation, and with a recombination
330 landscape conserved for a longer period of time (Cutter and Payseur 2013). Variation in the
331 relative impact of gene flow versus linked selection among species with comparable rates of
332 gene flow may therefore be explained by different mechanisms regulating recombination (e.g.
333 Youds and Boulton 2011; Serrentino and Borde 2012; Baudat et al. 2013), or differences in
334 turnover rates of karyotypes affecting the stability of recombination cold spots, such as
335 centromeres, the latter of which likely co-localize with differentiation islands within
336 chromosomes of *Ficedula* flycatchers (Ellegren et al. 2012). The strong recombination rate
337 variation observed in birds (Groenen et al. 2009; Backström et al. 2010; Kawakami et al. 2014b),
338 and their stable karyotypes (Ellegren 2010) may therefore be highly conductive to patterns such
339 as the ones reported here from *Ficedula* flycatchers.

340 Finally, our results call for utter caution with the interpretation of genome scans for
341 adaptively evolving genome regions that are based on differentiation (F_{ST} and related measures,
342 such as d_A and d_f) or diversity and that do not take into account recombination rate variation or
343 its effects on intraspecific genetic diversity. Such scans are likely to identify recombination-

mediated elevations of differentiation not necessarily attributable to selective sweeps. Likewise, our findings argue that the presence of high differentiation in the same genome regions of closely related taxa should not be mistaken as evidence for parallel evolution. Rather, this pattern is a direct prediction of linked selection in low-recombination regions of taxa amongst which the recombination landscape is conserved. The effects of linked selection in heterogeneous recombination landscapes should therefore be taken into account to formulate appropriate null models to reliably identify genome regions involved in speciation or adaptation.

Methods

Sampling

79 collared flycatchers and 79 pied flycatchers were sampled from four localities (20 birds from each population, 10 males and 10 females, except in two cases with 19 individuals; **Supplemental Table S1**), with varying degree of geographic overlap with the range of the other species. Collared flycatchers were sampled in Italy, Hungary, Czech Republic, and on the Baltic Sea island Öland (referred to as 'Baltic') in Sweden. Pied flycatchers were sampled in Spain, Sweden (mainland), Czech Republic, and on Öland ('Baltic') in Sweden. In addition, 20 Atlas flycatchers (*F. speculigera*; 14 males, 6 females) were sampled in the Moroccan Atlas Mountains and 20 semicollared flycatchers (*F. semitorquata*; 9 males, 11 females) were sampled in Bulgaria. A red-breasted flycatcher (*F. parva*) from Sweden and a snowy-browed flycatcher (*F. hyperythra*) from Indonesia were included as outgroups to polarize polymorphisms. According to mitochondrial cytochrome b sequence divergence, the divergence time of red-breasted flycatcher and of snowy-browed flycatcher to the black-and-white flycatcher complex dates back to approximately 4-5 Mya and 6 Mya, respectively.

DNA extraction, genome sequencing, and SNP-array genotyping

Samples consisted of blood stored in 96% ethanol (Spain, Czech Republic, Baltic pied flycatcher) or Queen's Lysis buffer (Italy, Atlas flycatcher, semicollared flycatcher) stored at 4°C (Atlas and semicollared flycatcher) or -20°C (all others). Tissues from Sweden, Baltic collared flycatcher, and red-breasted flycatcher consisted of muscle stored in 96% ethanol. The tissue of snowy-browed flycatcher consisted of muscle stored in salt buffer containing DMSO.

DNA was prepared using Qiagen's Blood and Tissue Kit following the manufacturer's instruction (Qiagen, Sollentuna, Sweden), including RNA-digestion, and avoiding strong vortexing in order to limit sharing of the DNA. Whole genome re-sequencing was performed with Illumina paired-end sequencing technology, using a HiSeq 2000 instrument at the SNP&Seq Technology Platform of Uppsala University. Individually tagged libraries with insert sizes of approximately 450 bp were constructed and sequenced from both ends using 100 cycles.

Twelve individuals from each collared and pied flycatcher population (except 6 for the Baltic pied population) were genotyped on a custom 50K Illumina iSelect SNP array according to protocols described elsewhere (Kawakami et al. 2014a), for use as training data set for VQSR.

Data preparation

All raw sequencing reads were mapped to a repeat-masked version of the collared flycatcher genome assembly FicAlb1.5 (Kawakami et al. 2014b) using BWA 0.7.4 (Li and Durbin 2009) with a soft-clipping base-quality threshold of 5 to avoid low-quality bases in alignments. 94.2 ± 3.3 % of the reads mapped to the reference genome on average, and reads from all four focal species mapped equally well (**Supplemental Table S2**). Mapping success was reduced to 83.6% and 72.6% in the near (*F. parva*) and far (*F. hyperythra*) outgroup, respectively. Alignment quality was enhanced by local realignment with GATK (McKenna et al. 2010; DePristo et al. 2011). Duplicates were marked at the library level using Picard (http://broadinstitute.github.io/picard/). Final sequencing coverage (excluding duplicates) was

393 14.7 x per individual (min. 5.0 x, max. 26.7 x, **Supplemental Table S2**) when considering the
394 total assembly length (1.1 Gb).
395 All subsequent steps were performed for each population separately to account for
396 population structure in the process of SNP identification. Base quality score recalibration
397 (BQSR) requires knowledge of variant sites. However, no extensive catalogs of such sites were
398 available for flycatchers. We therefore applied an approach to identify a reliable set of variant
399 sites that iteratively calls variants and uses a high-quality variant set from these for BQSR. The
400 initial round of variant calling using the original bam files was performed using
401 UnifiedGenotyperin GATK, SAMtools (Li et al. 2009), and FreeBayes (Garrison and Marth 2012)
402 using default settings. Variant sites overlapping between all three methods were extracted.
403 From these sites, SNPs for which at least one non-reference homozygote individual was
404 identified were included in the set of sites to be input into BQSR as variant sites. A second round
405 of variant calling was then run, using GATK only. As a test, we performed a second round of
406 BQSR for one population, and since variant calls were consistent between the second and third
407 rounds of calling, we considered the procedure to have converged and thus refrained from
408 employing a third round. For analyses that required genotypes, variant quality-score
409 recalibration was performed. Following BQSR, the SNP calls were analyzed and the highest
410 scoring 20% of SNPs and 34,786 variant sites confirmed by SNP-tying on the SNP-array were
411 used as training data set for VQSR in GATK. SNPs contained in the 90% and 99% tranches were
412 retained, resulting in ~11-22 million SNPs per population/species (**Supplemental Table S2**).
413 Finally, a catalogue of all sites variable within and between all populations and species was
414 assembled using GATK's CombineVariants, and all sites (73,366,635 sites when including
415 outgroups, 50,005,568 without outgroups) were genotyped using UnifiedGenotyper. Genotyping
416 was performed for each population separately in order to account for allele frequency
417 stratification. Final analyses for each population only considered sites with data for all
418 individuals in that population.

Repeat annotation

After the recent release of a second-generation collared flycatcher genome assembly (Kawakami et al. 2014b), we updated the repeat annotation in order to optimize the exclusion of repeat-derived regions. We screened the FicAlb1.5 assembly for flycatcher-specific repeats by employing RepeatModeler (version 1.0.5; http://www.repeatmasker.org/RepeatModeler.html), a *de-novo* repeat identification and modelling package that uses RMBlast (http://www.repeatmasker.org/RMBlast.html) and integrates the programs RECON version 1.07 (Bao and Eddy 2002), RepeatScout version 1.0.5 (Price et al. 2005), and Tandem repeats finder version 4.0.4 (Benson 1999). We manually curated the resulting repeat candidate library following standard procedures (Lavoie et al. 2013). Briefly, curation was done via BLASTn (Altschul et al. 1990) searching for long terminal repeat (LTR) retrotransposon-like repeat candidates against FicAlb1.5 and extraction of up to 50 of the best hits along with 1 kb of flanking sequence, respectively. For each of these LTR repeat candidates, we aligned the consensus sequence with its BLASTn hits using MAFFT (Katoh and Toh 2008) and subsequently generated a new, manually inspected consensus sequence. Each consensus sequence was considered to be complete only if it was flanked by single-copy sequence at its 5' and 3' ends within the alignment. After combining the flycatcher repeat library (containing 30 LTR subfamilies with complete consensus sequences and 249 other, possibly incomplete repeat consensus sequences) with previously known avian repeat elements (mainly from chicken and zebra finch) available in Repbase (http://www.girinst.org/repbase/index.html), we used this custom repeat library to annotate and mask the collared flycatcher genome assembly via RepeatMasker (version 3.2.9; http://www.repeatmasker.org/ RMDownload.html). The masked version was used in all subsequent analyses.

Estimation and analysis of population genetic parameters

Population genetic parameters were estimated for non-overlapping 200 kb windows along the genome, as 200 kb was the highest resolution for which pedigree-based recombination rates could be estimated. Population genetic inference was based on genotype likelihoods. ANGSD (Korneliussen et al. 2014) was used to estimate allele frequency likelihoods, obtain a maximum likelihood estimate of the unfolded site frequency spectrum (SFS) (Nielsen et al. 2012), and estimate diversity and SFS statistics including Tajima's *D*, and Fay and Wu's *H* (Fay and Wu 2000). Only sites with a minimal mapping quality of 1 and minimal quality score of 20 were considered. The ancestral sequence was reconstructed using genotypes from the two outgroup species based on parsimony. Genetic differentiation (F_{ST}) based on genotype likelihoods was estimated based on the two-dimensional SFS using ngsTools (Fumagalli et al. 2014). Sequence divergence (d_{xy}) was estimated from minor allele frequencies estimated in ANGSD. Additionally, to obtain an estimate of sequence divergence corrected for mutation rate, relative node depth (RND, Feder et al. 2005), was estimated by dividing d_{xy} between focal species with the average of d_{xy} between each focal species and snowy-browed flycatcher. To obtain the latter, d_{xy} was estimated based on genotype data from the collared (Italy), pied (Spain), Atlas, and semicollared flycatcher individual (male) with highest mean sequencing coverage. In order to obtain an estimator for the amount of allele frequency changes specific to each population, we estimated the population branch statistic (*PBS*) (Shriver et al. 2004; Yi et al. 2010) for collared, pied, Atlas, and semicollared flycatchers. Final analyses included 4,961 200 kb windows containing at least 100,000 sites passing the filtering criteria in ANGSD.

To study differentiation across longer evolutionary time scales, we estimated F_{ST} between the four black-and-white flycatchers and the two outgroups. Since only single outgroup individuals were sequenced, only one individual of each species was used for this analysis (the male with highest mean sequencing coverage) and only loci with data available for all individuals in the population sample were included. Weir and Cockerham's unbiased estimator of F_{ST} (Weir and Cockerham 1984) was estimated based on genotypes using Yangs's (Yang 1998)

hierarchical estimation procedure implemented in the HIERFSTAT package (Goudet 2005) in R. Multi-SNP F_{ST}-estimates for the 200 kb windows were obtained as ratio of averages of variance components.

Recombination rate estimates in cM/Mb for 200 kb windows were taken from Kawakami et al (2014b).

To investigate the relationship between the population genetic parameters and genomic features we performed principal component analyses (PCA) of the four π estimates, six d_{xy} estimates, six *RND* estimates, and four *PBS* estimates and used the respective first axis (PC1) as representative for the common variation in diversity, sequence divergence, and differentiation, respectively. We performed multiple linear regression analysis with either PC1(π), PC1(d_{xy}), PC1(*RND*), or PC1(*PBS*) as response variable and recombination rate, exon density, and their interaction term as candidate explanatory variables. To reduce the skewedness in their distribution we transformed PC1(*PBS*) by log-transformation to base 10, recombination rate by log-transformation to base 10, after adding a constant of 1 to keep zero rate values, and exon density by square root transformation. The same analysis was done using F_{ST} instead of *PBS*, providing highly similar results (data not shown). To investigate the correlation of the interspecific differentiation landscape with the differentiation landscapes within collared and within pied flycatcher, we preformed linear regressions of PC1(*PBS*) from interspecific comparisons obtained as described above against *PBS* estimates obtained from within-species comparisons, and estimated the overlap of interspecific differentiation islands with windows attributed to differentiation islands within any of the six within-species comparisons within collared or pied flycatcher respectively. All statistical analyses were performed using autosomal windows exclusively.

To test for each differentiation island whether it had significantly reduced Fay and Wu's *H* and Tajima's *D*, we compared the mean value in each island against the background distribution of these parameters. The latter was obtained by excluding all windows attributed to

differentiation islands in any species. Significance of the tests was estimated after sequential Bonferroni correction.

Unless otherwise stated, parameter estimates of all figures are shown smoothed using a Savitzky-Golay filter with a cubic regression within five consecutive windows.

Inference of differentiation islands

In order to detect autosomal 200 kb windows of high differentiation, a null distribution of differentiation for each pairwise comparison was obtained by randomizing autosomal SNPs across SNP positions, while maintaining the number of windows and the distribution of the number of SNPs per window. From this, *PBS* cut-off values were obtained by calculating *PBS* from the upper 0.5% percentile of the null distribution. Observed *PBS*-values were smoothed using a Savitzky-Golay smoothing filter in R with a cubic regression within 15 consecutive windows. Windows where then assigned to differentiation islands if their smoothed lineage-specific F_{ST}-value exceeded the respective cut-off. Interspecific differentiation islands were defined as windows situated in differentiation islands in all interspecific comparisons.

Estimation of shared derived variation

To infer gene flow among the four flycatcher species, we applied a test based on Patterson's *D*-statistic (Kulathinal et al. 2009; Green et al. 2010; Durand et al. 2011) that uses an asymmetric four-species tree setting (also known as 'ABBA-BABA' test). In such a framework, an excess of shared derived variation between the first outgroup species and one of the inner species ('ABBA') versus the other inner species ('BABA') is interpreted gene flow into the species showing the excess. To assess genome-wide patterns of introgression in species with population samples, we calculated the frequency of 'ABBA' and 'BABA' patterns using the formula $(1\text{-}p_1)p_2p_3$ and $p_1(1\text{-}p_2)p_3$, respectively, with p_i being the estimated derived allele frequencies for the three inner species obtained from ANGSD (Green et al. 2010; Durand et al. 2011). The fourth species is used to determine the ancestral state of SNPs and is not required in our case, since we already

polarized variant sites with data from *F. parva* and *F. hyperythra*. We then summed the frequencies of the 'ABBA' and 'BABA' site patterns over contiguous 5-Mb windows (n = 197) and obtained standard errors by applying a block-jackknife procedure (Green et al. 2010). We chose the length of the blocks to be much larger than the occurrence of linkage disequilibrium in flycatchers, even in low-recombination regions, therefore addressing the problem of non-independence among sites (Kawakami et al. 2014a). We then calculated the mean and variance of the *D*-statistic among the 197 leave-one-out replicates by weighting by the number of polymorphic sites within each 5-Mb window. This way, we obtained approximately normally distributed standard errors (Efron 1981; Reich et al. 2009), from which we calculated z-scores against an expected value of the *D*-statistic of zero. We then obtained the corresponding two-tailed p-values from the cumulative function of the standard normal distribution.

Genealogical sorting

To assess genome-wide variation in lineage sorting within each of the four focal species, we calculated the genealogical sorting index (*gsi*) (Cummings et al. 2008) for each species in 200 kb windows. The *gsi* statistic quantifies the degree of clustering of haplotypes in a phylogenetic tree, with a value of one indicating that a species forms a monophyletic group and a value of zero indicating a random distribution of haplotypes in the tree. To minimize recombination within and linkage between loci, we calculated *gsi* statistics for 10 kb windows with a spacing of 40 kb between subsequent windows and then averaged over four windows to obtain statistics for the same 200 kb windows as used in the other population genomic analyses. For each 10 kb window, we first performed statistical phasing with fastPHASE v1.4.0 (Scheet and Stephens 2006), using a fixed number of ten clusters and otherwise default settings. We phased all 198 individuals of the four flycatcher species together, but provided subpopulation structure to the program based on the species designation. We coded heterozygous positions with less than 80% posterior phasing support as missing data. Furthermore, we randomly selected one haplotype per individual for further analysis to minimize the effect of phasing error. Due to unequal

549 sampling sizes in the four species, we subsampled collared and pied flycatchers to 20 individuals
550 each after phasing by selecting the five individuals per population with the highest sequence
551 coverage for each window. For the two outgroup individuals, we produced haplotypes by
552 selecting one of the two alleles for each heterozygous position at random. Using an alignment of
553 82 haploid sequences, we built a phylogenetic tree applying the GTRGAMMA substitution model
554 in RAxML v8.0.20 (Stamatakis 2014). We rooted the resulting tree with the *F. hyperythra*
555 outgroup using Newick Utilities v1.6 (Junier and Zdobnov 2010), and calculated *gsi* statistics on
556 the rooted tree with a custom-made Perl script using the BioPerl Tree module (Stajich et al.
557 2002).

558 Species tree reconstruction

559 To infer the species tree of the four closely related flycatcher species and the two outgroups, we
560 analysed a large number of gene trees in a multispecies coalescent framework implemented in
561 the software MP-EST v1.4 (Liu et al. 2010). For this, we reused the same set of gene trees from
562 10 kb windows as inferred for the *gsi* statistic, but excluded windows on the Z chromosome,
563 resulting in a total of 17,504 gene trees. We then performed ten independent runs of MP-EST
564 and retained the species tree with the highest pseudo-likelihood (**Supplemental Fig. S1**). To get
565 bootstrap support values for each clade, we applied a gene tree resampling approach. We
566 performed 1,000 bootstrap replicates by resampling with replacement 17,504 trees from the set
567 of gene trees. Each bootstrap replicate was then used to run MP-EST as for the original species
568 tree inference. The set of species trees from the bootstrap replicates were summarized with
569 Newick Utilities (Junier and Zdobnov 2010), and the resulting bootstrap support values were
570 added to the species tree inferred from the original data.
571
572
573

Data access

Re-sequencing data from this study have been submitted to the EMBL-EBI European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under accession number PRJEB7359. Custom Perl and R scripts used in this study are available as Supplemental Scripts.

Acknowledgements

This research was supported by the European Research Council (AdG 249976), the Knut and Alice Wallenberg Foundation (Wallenberg Scholar Grant) and the Swedish Research Council (2010-5650, all to HE), the Swiss National Science Foundation (PBLAP3-134299 and PBLAP3_140171 to RB), and the Czech Science Foundation (P302/12/1207 to SB). Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, which is supported by Uppsala University, Uppsala University Hospital, Science for Life Laboratory Uppsala and the Swedish Research Council (80576801 and 70374401). We thank Elodie Kuhnert, Jochen Wolf, Egidio Fulco, Stefano Sarrocco, and Fernando Spina, and the Instituto Superiore per la Protezione e la Ricerca Ambientale for support with fieldwork. We are grateful to Matteo Fumagalli and Thorfinn Korneliussen for assistance with ngsTools and ANGSD software, respectively. The Natural History Museum of Stockholm and the Zoological Museum of the University of Copenhagen kindly contributed tissue sample of red-breasted flycatcher (*Ficedula parva*; accession NRM996601) and snowy-browed flycatcher (*F. hyperythra*; accession 148317), respectively. The Natural History Museum in Oslo stored DNA samples. We thank Chris Jiggins and two additional reviewers for their constructive comments on an earlier version of the manuscript.

Author contributions

RB and HE conceived the study. RB, SB, LZG, SH, JM, AQ, MR, SAS, GPS, and JT performed field-work. RB and PIO performed bioinformatics analyses. AS was responsible for the repeat

598 annotation. LS updated the reference assembly. RB, AN, TK, CFM and LD conducted population
599 genomic analyses. RB and HE wrote the manuscript with input from the other authors, and
600 prepared the Supplemental Materials with contributions from AN and AS.

601 Disclosure declaration

602 The authors declare no conflicts of interest.

603 Figure legends

604 **Figure 1.** A recurrently evolving genomic landscape of differentiation across the speciation
605 continuum in *Ficedula* flycatchers. **(a)** Species' neighbour-joining tree based on mean genome-
606 wide net sequence divergence (d_A). The same species tree topology was inferred with 100%
607 bootstrap support from the distribution of gene trees under the multispecies coalescent
608 (**Supplemental Fig. S1**). **(b)** Map showing the locations of population sampling and
609 approximate species ranges **(c)** Population genomic parameters along an example chromosome
610 (Chromosome 4A; see **Supplemental Fig. S2, S4** for all chromosomes). Color codes for specific-
611 specific parameters: blue, collared; green, pied; orange, Atlas; red, semicollared. Color codes for
612 d_{xy}: green, collared-pied; light blue, collared-Atlas; blue, collared-semicollared; orange, pied-
613 Atlas; red, pied-semicollared; black, Atlas-semicollared. For differentiation within species,
614 comparisons with the Italian (collared) and Spanish (pied) populations are shown. Color codes
615 for F_{ST} within collared flycatchers: cyan, Italy-Hungary; light blue, Italy-Czech Republic; dark
616 blue, Italy-Baltic. Color codes for F_{ST} within pied flycatchers: light green, Spain-Sweden; green,
617 Spain-Czech Republic; dark green, Spain-Baltic. **(d)** Distributions of differentiation (F_{ST}) from
618 collared flycatcher along the speciation continuum. Distributions are given separately for three
619 autosomal recombination percentiles (33%, 33-66%, 66-100%) corresponding to high (>3.4
620 cM/Mb, blue), intermediate (1.3-3.4 cM/Mb, orange), and low recombination rate (0-1.3 cM/Mb,
621 red), and the Z chromosome (green). Geographically close within-species comparison: Italy-

Hungary. Comparisons within species include the geographically close Italian and Hungarian populations [within (close)], and the geographically distant Italian and Baltic populations [within (far)]. Geographically far within-species comparison: Italy-Baltic. **(e)** Differentiation from collared flycatcher along an example chromosome (Chromosome 11; see **Supplemental Fig. S3** for all chromosomes). Color codes for between-species comparisons: green, pied; orange, Atlas; red, semicollared; dark red, red-breasted; black, snowy-browed flycatcher. Color codes for within-species comparisons: cyan, Italy-Hungary; blue, Italy-Baltic. Flycatcher artwork in panel (a) courtesy of Dan Zetterström.

Figure 2. Differentiation in independent lineages and its correlation with recombination rate. (**a**) Correlation of F_{ST} in 200 kb windows among two phylogenetically independent comparisons (Pearson's correlation: R=0.804, t=96.6, $p<10^{-15}$). (**b**) Relationship between differentiation and recombination rate (r, cM/Mb). Differentiation is expressed as the first axis (PC1) from a PCA on lineage-specific F_{ST} (*PBS*; linear regression, t=-40.7, $p<10^{-15}$, $R^2=0.266$).

Figure 3. Site-frequency spectrum statistics across an example chromosome (Chromosome 10). Color codes: blue, collared; green, pied; orange, Atlas; red, semicollared flycatcher. A signal of selection as indicated by negative Tajima's *D* is seen in the centrally located island in all species, while evidence for positive selection as indicated by negative Fay and Wu's *H*, i.e. an excess of high-frequency derived variants, is seen in only one species (pied flycatcher). See **Supplemental Fig. 2** for all chromosomes.

References

Alatalo RV, Eriksson D, Gustafsson L, Lundberg A. 1990. Hybridization between pied and collared flycatchers—sexual selection and speciation theory. *J Evol Biol* **3**: 375-389.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Andrew RL, Rieseberg LH. 2013. Divergence is focused on few genomic regions early in speciation: Incipient speciation in sunflower ecotypes. *Evolution* **67**: 2468-2482.

Backström N, Forstmeier W, Schielzeth H, Mellenius H, Nam K, Bolund E, Webster MT, Öst T, Schneider M, Kempenaers B et al. 2010. The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res* **20**: 485-495.

Backström N, Sætre G-P, Ellegren H. 2013. Inferring the demographic history of European *Ficedula* flycatcher populations. *BMC Evol Biol* **13**: 2.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**: 1269-1276.

Bateson W. 1909. Heredity and variation in modern lights. In *Darwin and modern science*, (ed. AC Seward), pp. 85-101. Cambridge University Press, Cambridge.

Baudat F, Imai Y, de Massy B. 2013. Meiotic recombination in mammals: Localization and regulation. *Nat Rev Genet* **14**: 794-806.

Benson G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573-580.

Borge T, Lindroos K, Nádvorník P, Syvänen AC, Sætre GP. 2005. Amount of introgression in flycatcher hybrid zones reflects regional differences in pre and post-zygotic barriers to gene exchange. *J Evol Biol* **18**: 1416-1424.

Charlesworth B. 1998. Measures of divergence between populations and the effect of forces that reduce variability. *Mol Biol Evol* **15**: 538-543.

Charlesworth B, Campos JL. 2014. The relations between recombination rate and patterns of molecular variation and evolution in *Drosophila*. *Annu Rev Genet* **48**: 383-403.

Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics* **134**: 1289-1303.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol Ecol* **23**: 3133-3157.

Cummings MP, Neel MC, Shaw KL. 2008. A genealogical approach to quantifying lineage divergence. *Evolution* **62**: 2411-2422.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: Unifying the disparity among species. *Nat Rev Genet* **14**: 262-274.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.

Dobzhansky T. 1937. *Genetics and the origin of species*. Columbia University Press, New York.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. *Mol Biol Evol*.

Efron B. 1981. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika* **68**: 589-599.

Ellegren H. 2010. Evolutionary stasis: The stable chromosomes of birds. *Trends Ecol Evol* **25**: 283-291.

Ellegren H, Smeds L, Burri R, Olason PI, Backström N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**: 756-760.

Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. *Genome Res*.

Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405-1413.

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends Genet* **28**: 342-350.

696 Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. 2014. Genome-wide
697 congealing and rapid transitions across the speciation continuum during speciation with
698 gene flow. *J Hered* **105**: 810-820.
699 Feder JL, Xie X, Rull J, Velez S, Forbes A, Leung B, Dambroski H, Filchak KE, Aluja M. 2005. Mayr,
700 Dobzhansky, and Bush and the complexities of sympatric speciation in *Rhagoletis*. *Proc*
701 *Natl Acad Sci USA* **102**: 6573-6580.
702 Flaxman SM, Feder JL, Nosil P. 2013. Genetic hitchhiking and the dynamic buildup of genomic
703 divergence during speciation with gene flow. *Evolution* **67**: 2577-2591.
704 Fumagalli M, Vieira FG, Linderoth T, Nielsen R. 2014. ngsTools: Methods for population genetics
705 analyses from next-generation sequencing data. *Bioinformatics*.
706 Garrison E, Marth G. 2012. Haplotype-based variation detection from short-read sequencing.
707 *arXiv*: 1207.3907.
708 Goudet J. 2005. HIERFSTAT, a package for r to compute and test hierarchical f-statistics. *Mol Ecol*
709 *Notes* **5**: 184-186.
710 Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz
711 MH-Y et al. 2010. A draft sequence of the neandertal genome. *Science* **328**: 710-722.
712 Groenen MAM, Wahlberg P, Foglio M, Cheng HH, Megens H-J, Crooijmans RPMA, Besnier F,
713 Lathrop M, Muir WM, Wong GK-S et al. 2009. A high-density SNP-based linkage map of
714 the chicken genome reveals sequence features correlated with recombination rate.
715 *Genome Res* **19**: 510-519.
716 Harr B. 2006. Genomic islands of differentiation between house mouse subspecies. *Genome Res*
717 **16**: 730-737.
718 Hendry AP, Bolnick DI, Berner D, Peichel CL. 2009. Along the speciation continuum in
719 sticklebacks. *J Fish Biol* **75**: 2000-2036.
720 Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC,
721 White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks.
722 *Nature* **484**: 55-61.

723 Junier T, Zdobnov EM. 2010. The newick utilities: High-throughput phylogenetic tree processing
724 in the Unix shell. *Bioinformatics* **26**: 1669-1670.
725 Katoh K, Toh H. 2008. Recent developments in the mafft multiple sequence alignment program.
726 *Brief Bioinform* **9**: 286-298.
727 Kawakami T, Backström N, Burri R, Husby A, Olason P, Rice AM, Ålund M, Qvarnström A,
728 Ellegren H. 2014a. Estimation of linkage disequilibrium and interspecific gene flow in
729 *Ficedula* flycatchers by a newly developed 50k single-nucleotide polymorphism array.
730 *Mol Ecol Res* **14**: 1248-1260.
731 Kawakami T, Smeds L, Backström N, Husby A, Qvarnström A, Mugal CF, Olason P, Ellegren H.
732 2014b. A high-density linkage map enables a second-generation collared flycatcher
733 genome assembly and reveals the patterns of avian recombination rate variation and
734 chromosomal evolution. *Mol Ecol* **23**: 4035-4058.
735 Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of next generation sequencing
736 data. *BMC Bioinformatics* **15**: 356.
737 Kronforst Marcus R, Hansen Matthew EB, Crawford Nicholas G, Gallant Jason R, Zhang W,
738 Kulathinal Rob J, Kapan Durrell D, Mullen Sean P. 2013. Hybridization reveals the
739 evolving genomic architecture of speciation. *Cell Reports* **5**: 666-677.
740 Kulathinal RJ, Stevison LS, Noor MAF. 2009. The genomics of speciation in *Drosophila*: Diversity,
741 divergence, and introgression estimated using low-coverage genome sequencing. *PLoS*
742 *Genet* **5**: e1000550.
743 Lavoie C, Platt R, Novick P, Counterman B, Ray D. 2013. Transposable element evolution in
744 *Heliconius* suggests genome diversity within lepidoptera. *Mobile DNA* **4**: 21.
745 Lawniczak MKN, Emrich SJ, Holloway AK, Regier AP, Olson M, White B, Redmond S, Fulton L,
746 Appelbaum E, Godfrey J et al. 2010. Widespread divergence between incipient *Anopheles*
747 *gambiae* species revealed by whole genome sequences. *Science* **330**: 512-514.
748 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform.
749 *Bioinformatics* **25**: 1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Liu L, Yu L, Edwards S. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol Biol* **10**: 302.

Lundberg A, Alatalo RV. 1992. *The pied flycatcher*. T & AD Poyser Ltd.

Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, Blaxter M, Manica A, Mallet J, Jiggins CD. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res*.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Muller HJ. 1940. Bearing of the *Drosophila* work on systematics. In *The new systematics*, (ed. J Huxley). Clarendon Press, Oxford.

Nachman MW. 2002. Variation in recombination rate across the genome: Evidence and implications. *Curr Opin Genet Dev* **12**: 657-663.

Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: Theoretical predictions and empirical results from rabbits and mice. *Phil Trans R Soc B* **367**: 409-421.

Nadachowska-Brzyska K, Burri R, Olason PI, Kawakami T, Smeds L, Ellegren H. 2013. Demographic divergence history of pied flycatcher and collared flycatcher inferred from whole-genome re-sequencing data. *PLoS Genet* **9**: e1003942.

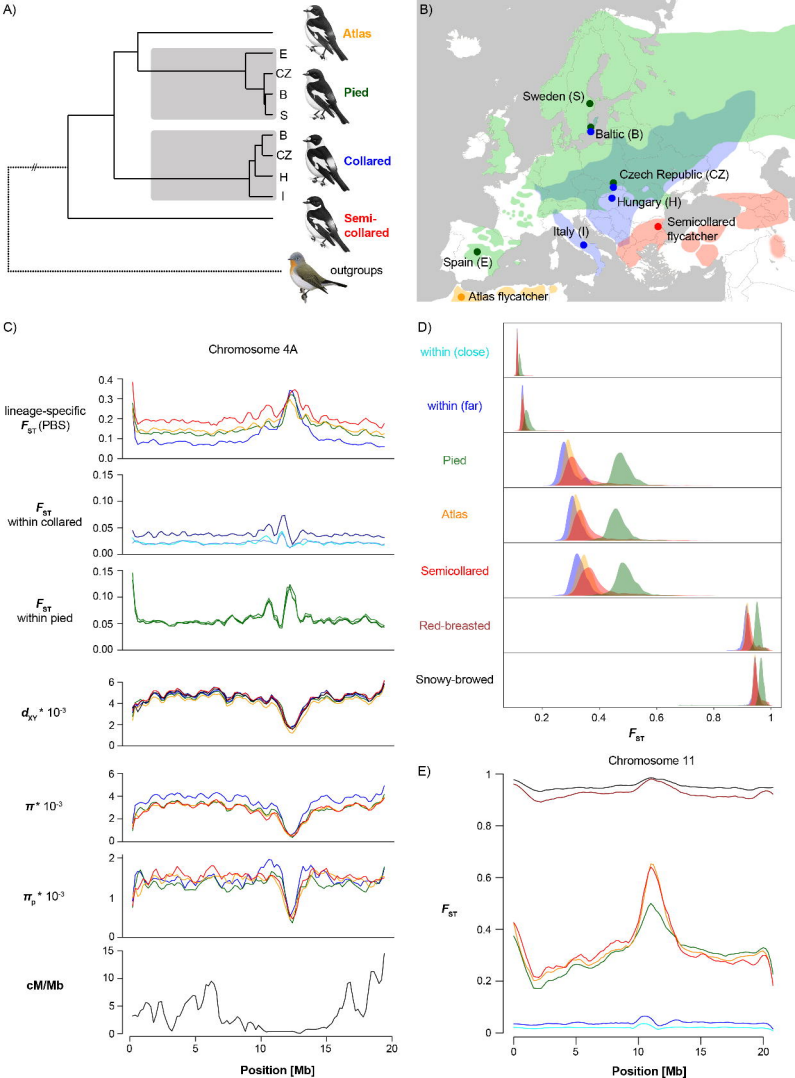
Nadeau NJ, Whibley A, Jones RT, Davey JW, Dasmahapatra KK, Baxter SW, Quail MA, Joron M, ffrench-Constant RH, Blaxter ML et al. 2012. Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil Trans R Soc B* **367**: 343-353.

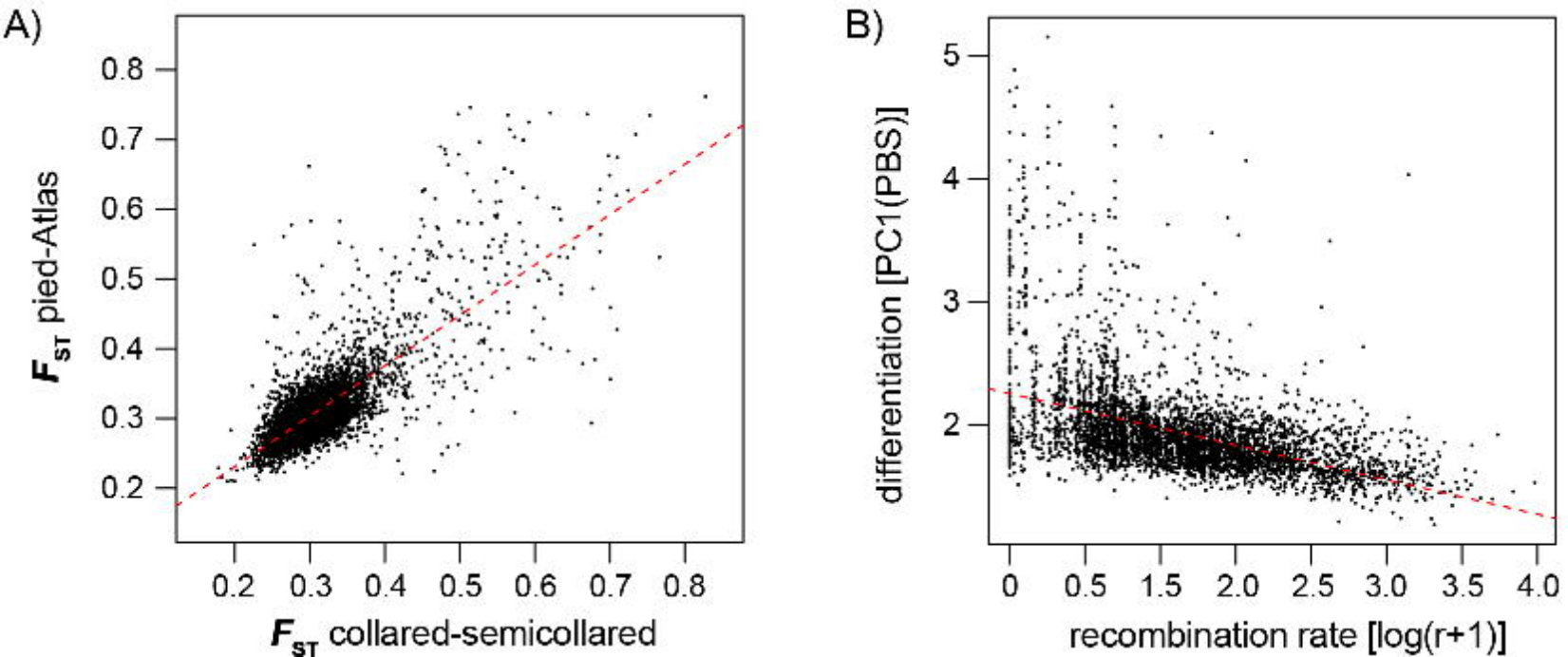
776 Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. in press. Resolving evolutionary
777 relationships in closely related species with whole-genome sequencing data. *Syst Biol*.
778 Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J. 2012. Snp calling, genotype calling, and
779 sample allele frequency estimation from new-generation sequencing data. *PLoS ONE* **7**:
780 e37558.
781 Noor MAF, Bennett SM. 2009. Islands of speciation or mirages in the desert? Examining the role
782 of restricted recombination in maintaining species. *Heredity* **103**: 439-444.
783 Nosil P. 2008. Speciation with gene flow could be common. *Mol Ecol* **17**: 2103-2106.
784 Nosil P, Egan SP, Funk DJ. 2008. Heterogeneous genomic differentiation between walking-stick
785 ecotypes: "Isolation by adaptation" and multiple roles for divergent selection. *Evolution*
786 **62**: 316-336.
787 Patterson NJ, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, Genschoreck T, Webster T, Reich
788 D. 2012. Ancient admixture in human history. *Genetics*.
789 Pennisi E. 2014. Disputed islands. *Science* **345**: 611-613.
790 Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes.
791 *Bioinformatics* **21**: i351.
792 Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing indian population
793 history. *Nature* **461**: 489-494.
794 Renaut S, Grassa CJ, Yeaman S, Moyers BT, Lai Z, Kane NC, Bowers JE, Burke JM, Rieseberg LH.
795 2013. Genomic islands of divergence are not affected by geography of speciation in
796 sunflowers. *Nat Commun* **4**: 1827.
797 Renaut S, Owens GL, Rieseberg LH. 2014. Shared selective pressure and local genomic landscape
798 lead to repeatable patterns of genomic divergence in sunflowers. *Mol Ecol* **23**: 311-324.
799 Roesti M, Hendry AP, Salzburger W, Berner D. 2012. Genome divergence during evolutionary
800 diversification as revealed in replicate lake–stream stickleback population pairs. *Mol*
801 *Ecol* **21**: 2852-2862.

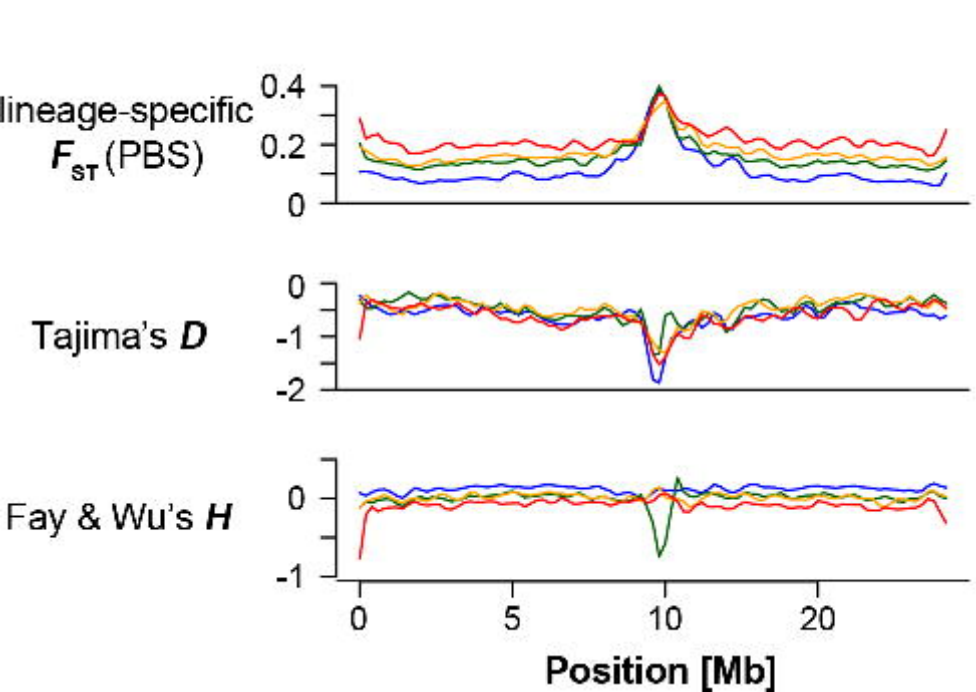
802 Sætre G-P, Sæther SA. 2010. Ecology and genetics of speciation in *Ficedula* flycatchers. *Mol Ecol*
803 **19**: 1091-1106.
804 Sætre GP, Král K, Bures S, Ims RA. 1999. Dynamics of a clinal hybrid zone and a comparison with
805 island hybrid zones of flycatchers (*Ficedula hypoleuca* and *F. albicollis*). *J Zool* **247**: 53-
806 64.
807 Scheet P, Stephens M. 2006. A fast and flexible statistical model for large-scale population
808 genotype data: Applications to inferring missing genotypes and haplotypic phase. *Am J*
809 *Hum Genet* **78**: 1611-1618.
810 Seehausen O, Butlin RK, Keller I, Wagner CE, Boughman JW, Hohenlohe PA, Peichel CL, Sætre G-
811 P, Bank C, Brannstrom A et al. 2014. Genomics and the origin of species. *Nat Rev Genet*
812 **15**: 176-192.
813 Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the
814 *Drosophila* genome? *PLoS Genet* **5**: e1000495.
815 Serrentino M-E, Borde V. 2012. The spatial regulation of meiotic recombination hotspots: Are all
816 dsb hotspots crossover hotspots? *Exp Cell Res* **318**: 1347-1352.
817 Shaw KL, Mullen SP. 2014. Speciation continuum. *J Hered* **105**: 741-742.
818 Shriver M, Kennedy G, Parra E, Lawson H, Sonpar V, Huang J, Akey J, Jones K. 2004. The genomic
819 distribution of population substructure in four populations using 8,525 autosomal SNPs.
820 *Human Genomics* **1**: 274 - 286.
821 Soria-Carrasco V, Gompert Z, Comeault AA, Farkas TE, Parchman TL, Johnston JS, Buerkle CA,
822 Feder JL, Bast J, Schwander T et al. 2014. Stick insect genomes reveal natural selection's
823 role in parallel speciation. *Science* **344**: 738-742.
824 Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I,
825 Lapp H et al. 2002. The BioPerl Toolkit: Perl modules for the life sciences. *Genome Res*
826 **12**: 1611-1618.
827 Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
828 phylogenies. *Bioinformatics* **30**: 1312-1313.

829 Tegelström H, Gelter HP. 1990. Haldane rule and sex biased gene flow between two hybridizing
830 flycatcher species (*Ficedula albicollis* and *F. hypoleuca*, Aves, Muscicapidae). *Evolution*
831 **44**: 2012-2021.
832 Turner TL, Hahn MW. 2010. Genomic islands of speciation or genomic islands and speciation?
833 *Mol Ecol* **19**: 848-850.
834 Turner TL, Hahn MW, Nuzhdin SV. 2005. Genomic islands of speciation in *Anopheles gambiae*.
835 *PLoS Biol* **3**: e285.
836 Via S, West J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological
837 speciation. *Mol Ecol* **17**: 4334-4345.
838 Weir BS, Cockerham CC. 1984. Estimating f-statistics for the analysis of population-structure.
839 *Evolution* **38**: 1358-1370.
840 White BJ, Cheng CD, Simard F, Costantini C, Besansky NJ. 2010. Genetic association of physically
841 unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Mol*
842 *Ecol* **19**: 925-939.
843 Wiley C, Qvarnström A, Andersson G, Borge T, Sætre G-P. 2009. Postzygotic isolation over
844 multiple generations of hybrid descendants in a natural hybrid zone: How well do single-
845 generation estimates reflect reproductive isolation? *Evolution* **63**: 1731-1739.
846 Yang RC. 1998. Estimating hierarchical f-statistics. *Evolution* **52**: 950-956.
847 Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, Xu X, Jiang H, Vinckenbosch N,
848 Korneliussen TS et al. 2010. Sequencing of 50 human exomes reveals adaptation to high
849 altitude. *Science* **329**: 75-78.
850 Youds JL, Boulton SJ. 2011. The choice in meiosis – defining the factors that influence crossover
851 or non-crossover formation. *J Cell Sci* **124**: 501-513.

852
853


A)
Atlas
Pied
Collared
Semi-
collared
outgroups
E
CZ
B
S
B
CZ
H
I
B)
Sweden (S)
Baltic (B)
Czech Republic (CZ)
Hungary (H)
Italy (I)
Spain (E)
Semicollared
flycatcher
Atlas flycatcher
C)
Chromosome 4A
lineage-specific
F_{ST} (PBS)
F_{ST}
within collared
F_{ST}
within pied
d_{XY} * 10^{-3}
π * 10^{-3}
π_p * 10^{-3}
cM/Mb
Position [Mb]
D)
within (close)
within (far)
Pied
Atlas
Semicollared
Red-breasted
Snowy-browed
F_{ST}
E)
Chromosome 11
F_{ST}
Position [Mb]


A)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
F_{ST} pied-Atlas
0.2
0.3
0.4
0.5
0.6
0.7
0.8
F_{ST} collared-semicollared
B)
5
4
3
2
differentiation [PC1(PBS)]
0
0.5
1.5
2.0
2.5
3.0
3.5
4.0
recombination rate [log(r+1)]


lineage-specific
F_{ST}(PBS)
0.4
0.2
0
Tajima's *D*
0
-1
-2
Fay & Wu's *H*
0
-1
0
5
10
20
Position [Mb]