MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 4757–4769 doi: 10.1111/mec.12842

# A role for migration-linked genes and genomic islands in divergence of a songbird

KRISTEN RUEGG,*† ERIC C. ANDERSON,‡§ JASON BOONE,¶ JAZZ POULS** and THOMAS B. SMITH*††

*Center for Tropical Research, Institute of the Environment and Sustainability, University of California, Los Angeles, La Kretz Hall, Suite 300, 619 Charles E. Young Dr. East, Los Angeles, CA 90095, USA, †Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA 95060, USA, ‡Southwest Fisheries Science Center, National Marine Fisheries Service, 110 Shaffer Road, Santa Cruz, CA 95060, USA, §Department of Applied Mathematics and Statistics, University of California, Santa Cruz, CA 95060, USA, ¶Floragenex, Inc., 44 West Broadway, Eugene, OR 97401, USA, **Pacific Collegiate School, 255 Swift Street, Santa Cruz, CA 95060, USA, ††Department of Ecology and Evolutionary Biology, University of California, Los Angles, CA 90095, USA

## Abstract

**Next-generation sequencing has made it possible to begin asking questions about the process of divergence at the level of the genome. For example, recently, there has been a debate around the role of 'genomic islands of divergence' (i.e. blocks of outlier loci) in facilitating the process of speciation-with-gene-flow. The Swainson's thrush, *Catharus ustulatus*, is a migratory songbird with two genetically distinct subspecies that differ in a number of traits known to be involved in reproductive isolation in birds (plumage coloration, song and migratory behaviour), despite contemporary gene flow along a secondary contact zone. Here, we use RAD-PE sequencing to test emerging hypotheses about the process of divergence at the level of the genome and identify genes and gene regions involved in differentiation in this migratory songbird. Our analyses revealed distinct genomic islands on 15 of the 23 chromosomes and an accelerated rate of divergence on the Z chromosome, one of the avian sex chromosomes. Further, an analysis of loci linked to traits known to be involved in reproductive isolation in songbirds showed that genes linked to migration are significantly more differentiated than expected by chance, but that these genes lie primarily outside the genomic islands. Overall, our analysis supports the idea that genes linked to migration play an important role in divergence in migratory songbirds, but we find no compelling evidence that the observed genomic islands are facilitating adaptive divergence in migratory behaviour.**

*Keywords*: birds, genomics, migration, sex chromosome, speciation

*Received 4 April 2014; revision received 13 June 2014; accepted 18 June 2014*

## Introduction

While our knowledge of the ecological and behavioural traits involved in the early stages of divergence has increased steadily (Coyne & Orr 2004; Price 2008), the tools for understanding how genomes diverge during the process of speciation in nonmodel organisms have emerged only in the last few years (Metzker 2010; Sousa & Hey 2013; Seehausen *et al.* 2014). As a result of advances in sequencing technology, we are now able to identify the genes and gene regions subject to selection and discern how they are embedded within the genomes of wild populations (Ellegren *et al.* 2012; Jones *et al.* 2012; Nosil & Feder 2012). At the same time, gene expression and association studies are increasing our understanding of the loci linked to traits involved in adaptation and this information can then be used in

Correspondence: Kristen Ruegg, Fax: (831) 459-5353; E-mail: kruegg@ucsc.edu

combination with genome scans to gain a more complete picture of the genomic landscape of divergence (Smadja *et al.* 2012).

One hypothesis that has arisen from recent genome-wide studies is the idea that some genomic regions, termed genomic islands of divergence, contain higher numbers of genes involved in speciation than other regions (Turner *et al.* 2005; Nosil *et al.* 2009; Via 2009; Feder *et al.* 2012a; Smadja *et al.* 2012). Speciation-with-gene-flow theories predict that blocks of highly diverged loci will arise as a result of strong divergent selection, genetic linkage and reduced recombination rates in some regions, while gene flow will break apart linkages between more weakly selected or neutral loci in other regions (Wu 2001; Turner *et al.* 2005; Feder *et al.* 2012a). Alternatively, some authors have hypothesized that in populations that have been isolated (allopatric), genetic divergence will be less heterogeneous across the genome as both drift and selection will further the process of differentiation unimpeded by the disruptive effects of gene flow (Feder *et al.* 2012a).

While speciation-with-gene-flow theoretical predictions are supported by some empirical data (Mallet *et al.* 2007; Nadeau *et al.* 2012), it has also been suggested that features of the underlying genomic architecture may be more important than the extent of gene flow in determining the size, location and number of genomic islands (Renaut *et al.* 2013; Yeaman 2013). In particular, regions of low recombination (chromosomal rearrangement sites, areas adjacent to centromeres, etc.) often show reduced intraspecific diversity and inflated interspecific divergence as a result of background selection against deleterious mutations and/or selective sweeps (Charlesworth *et al.* 1993; Charlesworth 1998; Noor & Bennett 2010). Thus, an island-like pattern of divergence may arise even in the absence of gene exchange. Presently, there is a need for more empirical data on the genomic landscape of divergence between populations with well-described ecological and evolutionary histories to further test predictions about the role of genomic islands in the process of speciation.

Another speciation theory that can be assessed with emerging genome-scale data is the idea that sex chromosomes play a disproportionately large role in the early stages of divergence (Dobzhansky 1974; Coyne 1985; Charlesworth *et al.* 1987; Ellegren 2009). The fast *X* phenomenon, or 'fast *X* on the *Z*', in reference to female heterogametic systems, as found in birds, predicts natural selection may be more efficient on sex chromosomes because beneficial mutations are directly exposed to positive selection in the heterogametic sex (Charlesworth *et al.* 1987; Begun *et al.* 2007; Mank *et al.* 2007). Alternatively, it has been hypothesized that sexual selection for exaggerated male traits in combination with linkage between preference and trait genes on sex chromosomes may accelerate the process of divergence on the sex chromosomes (Iyengar *et al.* 2002). While both of these hypotheses can be difficult to disentangle from the potential effect of genetic drift on sex chromosomes with smaller effective sizes (Mank *et al.* 2010), genome-scale data provide unprecedented opportunities to distinguish the signature of selection from other evolutionary forces on the sex chromosomes.

The Swainson's thrush (*Catharus ustulatus*) provides a model system in which to test emerging hypotheses about the process of divergence at the level of the genome because there is a good deal known about its ecology and evolutionary history. Within this species, there are two genetically distinct subspecies, a coastal form, *C. u. ustulatus,* that migrates along the Pacific Coast of North America, winters in Central America, and has reddish-coloured plumage and an inland form, *Catharus u. swainsoni,* that migrates along an eastern route, winters primarily in South America and has olive-coloured plumage (Fig. 1; Evans Mack & Yong 2000; Ruegg & Smith 2002; Ruegg *et al.* 2006b). Coales-

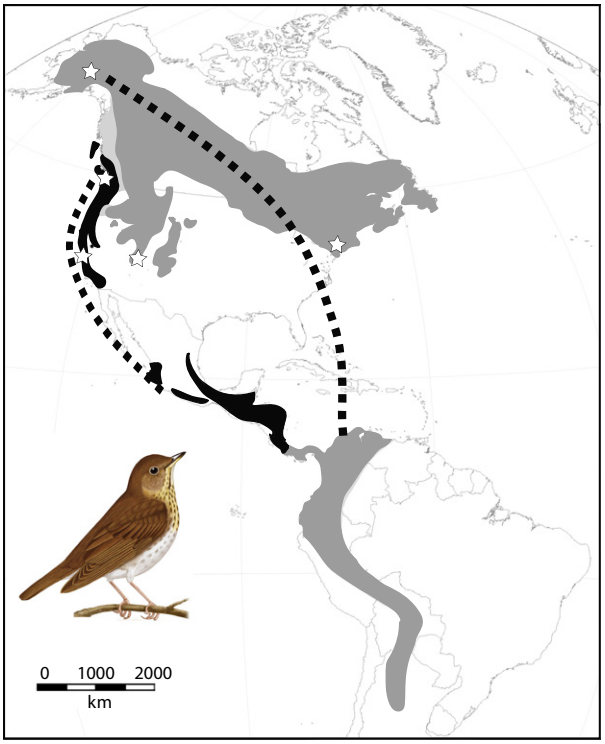


**Fig. 1** Range map and associated migratory pathways of coastal (black) and inland (grey) forms of the Swainson's thrush. Light grey indicates regions of potential hybridization, and the stars indicate regions where samples were collected for genetic analysis ($n = 5$ for each population). Swainson's thrush image was provided courtesy of the Mitch Waite Group.

cent analyses and ecological models suggest that the two forms probably diverged in eastern and western refugia sometime during the last glacial maximum and have since formed a zone of secondary contact in the Coast Mountains of British Columbia where gene flow is limited (Ruegg *et al.* 2006a; Ruegg 2008). The reason for the barrier to gene flow remains unclear, but the two forms differ in a number of traits known in birds to influence reproductive isolation (Price 2008): differences in arrival times on the breeding grounds at the centre of the hybrid zone, plumage coloration and song (Ruegg *et al.* 2006b, 2012; Ruegg 2008).

To investigate the role of genomic islands and sex chromosomes in the process of speciation, we combined RAD-PE sequencing with an analysis of genes linked to traits which we have an a priori reason to believe may be important to adaptive divergence in the Swainson's thrush (plumage coloration, song and migratory behaviour). Specifically, we ask the following questions: (i) What is the genomic landscape of divergence between subspecies experiencing limited gene flow along a putative zone of secondary contact? (ii) What is the role of genetic drift in generating the observed patterns of divergence? (iii) Are loci linked to traits known to be involved in reproductive isolation in birds, such as song, plumage coloration and migration more divergent than expected by chance? (iv) If yes, are these loci located within or outside the major genomic islands of divergence? Through the process of addressing the above questions, we also assess the utility of RAD-PE sequencing of individually barcoded samples for population genomic studies in taxa with limited genomic resources.

## Materials and methods

### *RAD-PE sequencing and bioinformatics*

We used restriction site-associated DNA paired-end (RAD-PE) sequencing to assess the genomic architecture of divergence in the Swainson's thrush. RAD-PE sequencing made it possible to build longer contigs (~300 bp) from short-read, 100-bp Illumina HiSeq2000 data to improve downstream gene mapping and ontology analysis (Etter *et al.* 2011). Twenty-five blood and tissue samples from five populations across the breeding range in North America were collected in adherence to university guidelines for Animal Care and Use at each institution (Table S1, Supporting information). Genomic DNA was purified using the Qiagen DNeasy Tissue Kit, and RAD-PE libraries were prepared at Floragenex, Inc. according to Baird *et al.* (2008). In short, 100 ng of genomic DNA from each individual was digested with 2 U of *SbfI-HF* enzyme (New England Biolabs, Beverly MA, USA) for 60 min at 37 °C. The reactions were then inactivated by holding at 65 °C for 20 min. The P1 adapter (a modified Illumina adapter, see Baird *et al.* 2008) was ligated to the products of the restriction reactions, and the 'barcoding' of the various samples was achieved with a set of index nucleotides within the P1 adapter sequence. One microlitre of 100 nM P1 adapter was then added to each sample with 60 U T4 DNA Ligase (Enzymatics, Inc.). Reactions were incubated at room temperature for 1 h and then heat-inactivated by holding at 65 °C for 10 min. The reactions were then pooled, and the products were randomly sheared to a mean size of 500 bp using a Bioruptor NGS (Diagenode). The material was electrophoresed through a 1.5% agarose gel, and DNA in the range of 200–700 bp was isolated using a MinElute Gel Extraction Kit (Qiagen). To remove overhangs, ssDNA ends were treated with 1 μL High-Concentration End-Repair Mix (Enzymatics, Inc.). The samples were purified by passing through a MinElute column (Qiagen), and 3′-adenine overhangs were added by the addition of 50 U Klenow (3′–5′ exo-) (Enzymatics) and 1 μL 10 mM dATP. Samples were then incubated at 37 °C for 30 min. Following repurification, 1 μL of 1 μM P2 adapter (a modified Illumina adapter, see Baird *et al.* 2008) was ligated with 600 U T4 DNA Ligase (Enzymatics, Inc.) for 1 h at room temperature. The samples were then purified as above and eluted in a volume of 15 μL. Following quantification using a Qubit fluorimeter (Invitrogen), 10 ng was taken as the template for a 100-μL PCR containing 50 μL Phusion Master Mix (NEB), 5 μL 10 μM P1 amplification primer and 5 μL 10 μM P2 amplification primer. The Phusion PCR settings followed standard protocols (NEB) over 18 cycles. Amplicons were then gel-purified, the size range 300–700 bp was excised from the gel, and its DNA content adjusted to 3.6 ng/μL.

Samples from each isolate were sequenced on an Illumina HiSeq2000 (Illumina, San Diego, CA, USA) using paired-end 100-bp sequence reads. Paired-end sequences from each sample were collected, separated by individual, stripped of barcodes, scrubbed of sequences that either appeared in very high read depths or that matched a database of putative contaminant sequences (ribosomal, prokaryotic, plastid), trimmed to 70 bp and filtered to include only those with a Phred score ≥10 (see Table S1, Supporting information for additional summary statistics on the sequencing run). The sample that by chance received the greatest number of reads passing the quality filter was used to create a reference set of RAD-PE contigs against which sequences from other samples were aligned. To create the reference, primary reads were clustered into unique RAD tags and the paired-end sequences associated with each RAD tag

were assembled de novo using VELVET (Zerbino & Birney 2008) into contigs ranging from 180–610 bp, with an average length of 300 bp (Fig. S1, Supporting information). Paired-end reads from the remaining samples were aligned to this reference, and SNPs were identified using the SAMTOOLS software with mpileup module under standard conditions (Li *et al.* 2009).

To narrow our data set to contigs containing high-quality SNPs we could confidently use to assess population structure, we discarded SNPs based upon the following stringent filtering steps: (i) putative SNPs with no variants remaining after data filtering or SNPs with more than two alleles; (ii) genotypes in individuals with a Phred score of <30; (iii) genotypes with <8 reads in a homozygote or <4 reads per allele in heterozygotes; and (iv) putative SNPs with genotypes in <10 out of the 15 samples from the *C. u. swainsoni* populations or <7 out of the 10 samples from the *C. u. ustulatus* populations. Only contigs containing at least one SNP that met all of our quality criteria were considered in the downstream analysis. To evaluate how divergence was arrayed across the genome, we mapped our contigs back to the closest, best annotated, songbird genome, the Zebra finch (*Taeniopygia guttata*; version 3.2.4; Warren *et al.* 2010) using BLAST+ (version 2.2.25). To avoid the possibility of erroneous matches, the data were filtered to include only contigs that matched with a single hit and an *E*-value < $10^{-40}$.

### *Population genomic analysis*

We used custom R scripts to calculate the following population genomic parameters: (i) the between-species population differentiation ($F_{ST}$); (ii) the density of fixed differences per base pair ($d_f$); (iii) the within-species nucleotide diversity ($\pi'$); and (iv) the between-species average number of pairwise differences ($d_{xy}$) $F_{ST}$ was estimated on a per-SNP basis, but the rest of the statistics were estimated for each contig in which a SNP was discovered. All scripts are available on Dryad (doi:10.5061/dryad.73gj4), but the methodology is briefly described below. In short, for each SNP, we calculated its contribution to the sample $\pi$ in each population as the probability that two randomly chosen sequences from the sample have different alleles at the SNP. For a SNP $i$ in population $x$, this is:

$$\pi^s_{x,i} = \frac{r_{i,x}a_{i,x}}{\binom{r_{i,x}+a_{i,x}}{2}}$$

where the superscript $s$ denotes 'at a SNP', and $r_{i,x}$ and $a_{i,x}$ are the number of reference and alternate alleles observed in the sample from population $x$ at SNP $i$. Using these values computed for two populations $x$ and $y$, we calculated $F_{ST}$ for each SNP, following Hohenlohe *et al.* (2010), as:

$$F_{ST} = 1 - \frac{\pi^s_{x,i}\binom{m}{2} + \pi^s_{y,i}\binom{n}{2}}{\pi^s_{all,i}\left(\binom{m}{2} + \binom{n}{2}\right)}$$

where $m = r_{i,x} + a_{i,x}$, $n = r_{i,y} + a_{i,y}$, and $\pi^s_{all,i}$ is the nucleotide diversity at SNP $i$ calculated when the samples from $x$ and $y$ are pooled together. For a contig $k$ of length $L_k$ having $S_k$ SNPs in it, we estimate the nucleotide diversity in population $x$ with:

$$\pi_{x,k} = \frac{1}{L_k}\sum_{i=1}^{S_k}\pi^s_{x,i},$$

and calculate the between-species sequence divergence as:

$$d_{xy,k} = \frac{1}{L_k}\sum_{i=1}^{S_k}\frac{r_{i,x}a_{i,y} + r_{i,y}a_{i,x}}{(r_{x,i}+a_{x,i})(r_{y,i}+a_{y,i})},$$

and the density of fixed SNPs, $d_f$, as the number of SNPs in the contig fixed for alternate alleles in the two populations, divided by $L_k$. As we used a stringent criterion to call SNPs, we might not have detected all the SNPs that were present in our data. This would tend to bias nucleotide diversity downward. Accordingly, we report the nucleotide diversity calculated only from those contigs that carried at least one SNP in our filtered data set. We refer to this as the zero-truncated nucleotide diversity, and we denote it as $\pi'$.

To determine how divergence is arrayed across the genome, we smoothed the parameters described above along the genome using a sliding window. Values for each statistic were calculated at closely spaced points by smoothing with a uniform kernel smoothing density (a 'box' density) of width 1 Mb centred on the point. To assess significance of genomic islands in a way that accounts for variation in marker density across the genome, we generated a sample of 25 000 smoothed lines under the null hypothesis of no spatial autocorrelation by permuting values amongst the locations of SNPs for each replicate and running our smoother on those values. Genomic islands were identified as any location where the observed smoothed values were greater than all 25 000 smoothed values simulated under the null hypothesis of no spatial autocorrelation. For parameters calculated on a per-contig basis, we weighted the smoothed average by contig length using code we implemented in C and called from R. For $N$ contigs on a chromosome, a statistic $y$ calculated for each contig $k$ was smoothed such that the smoothed value of $y$ at genome coordinate $z_j$ (in base pairs) on the chromosome

was as follows:

$$\hat{y}(z_j) = \frac{\sum_{k=1}^{N} y_k L_k Q(|z_j - G_k|)}{\sum_{k=1}^{N} y_k Q(|z_j - G_k|)}$$

where $G_k$ is the genome coordinate at the midpoint of contig $k$, and $Q()$ denotes the kernel density (in this case the uniform density over 1 Mb). Smoothing values of $F_{ST}$ proceeded similarly; however, as $F_{ST}$ is computed on a per-SNP basis, no weights (the $L_k$ terms) were necessary.

### *Simulations to test for influence of drift*

To determine the frequency distribution of genetic differentiation across loci and test whether or not the observed patterns of divergence could be explained by a simple model of drift, we conducted a series of simulations in which our observed distribution of $F_{ST}$ values was compared with those expected under a neutral island model. For each SNP in our data set, we simulated a single SNP from a neutral coalescent with migration between the subspecies parameterized by $4\ N_e m = 2$, 3, 4 or 5. Each SNP was simulated with the observed sample size (the number of sequences sampled from each subspecies varied amongst SNPs, because not all individuals were genotyped at every SNP) using a custom shell script making calls to the program makesamples (Hudson 2002). As SNPs in our data set were not ascertained in a separate sample, we were able to do this simply using makesample's –s option. $F_{ST}$ was calculated for every SNP in the simulated data set, and the distribution of $F_{ST}$ values from the simulations was visualized and compared to the observed distribution from the real data.

### *Gene ontology and candidate gene analysis*

To examine whether genomic islands and genes containing fixed differences were enriched for any particular gene category, we conducted an analysis of gene ontology terms (GO terms). All *T. guttata* genes (known, novel and putative) were downloaded from the BIOMART web portal (http://www.biomart.org/), and genes labelled 'protein_coding' were retained for the analysis (removing pseudogenes and snoRNAs). Genes with contigs mapping to any region of the gene (intron or exon) and without overlapping genome coordinates with proximate genes were retained for the analysis. SNPs were coded as being within an island of divergence if any sequence from the SNP's contig fell within an island. Enrichment analysis was performed using the tools available on the Zebra finch GO analysis website (http://bioinformatics.iah.ac.uk/tools/GOfinch; Wu & Watson 2009). Genes within islands and genes containing SNPs with $F_{ST} = 1.0$ were treated as the test gene list, while all genes sequenced within the context of this study were treated as the population gene list. Significance was determined using a Fisher's exact test with Benjamini and Hochberg corrections for multiple comparisons.

To identify genes linked to migration, pigmentation and song in birds, we compiled a list of candidate genes for each trait based upon a literature review of genes linked to the traits of interest in gene expression and association studies (Table S4; Appendix S1, Supporting information) and compared patterns of divergence in each of the genes for which we had RAD-seq data. To determine whether genes linked to any of the three traits (migration, song or pigmentation) were more divergent than expected by chance, we used a rank-based test as follows: consider $C$ candidate genes, $i = 1,\ldots,C$, in a category (e.g. migration or colour). Let candidate gene $i$ carry $X_i$ sequenced SNPs, amongst which the maximum $F_{ST}$ is $F_{max,i}$. We let $R_i$ be the fraction of all genes in our data set that also have $X_i$ SNPs, amongst which the highest $F_{ST}$ is greater than or equal to $F_{max,i}$. Our test statistic is then the sum over $i$ (i.e. over the C candidate genes in the category) of $\log(R_i)$. Use of the maximum $F_{ST}$ in a gene rather than the average $F_{ST}$ eliminates the confounding effects of varying levels of low-frequency singleton SNPs. We evaluate the significance of the test statistic by comparison with the distribution of values obtained by sampling sets of $C$ genes randomly from those in the autosomes (or the Z chromosome, should candidate gene $i$ be found on the Z) and computing the test statistic for each set. Genes that were more divergent than expected by chance were mapped using coordinates from the Zebra finch genome to determine their proximity to identified genomic islands of divergence.

## Results

### *RAD-PE sequencing and bioinformatics*

Overall, RAD-PE sequencing resulted in 132 645 contigs, containing 574 257 SNPs (Table S2, Supporting information). The average coverage across all contigs within an individual was ~29*X*, and the average contig length was ~300 bp (Table S1, Fig. S1, Supporting information). Subsequent SNP quality filtering reduced the total number of contigs by over one half, yielding a total of 64 513 contigs, containing 360 632 SNPs and of the 64 513 high-quality contigs that passed our filters, a total of 33 841 contigs, containing 154 123 SNPs, mapped to the Zebra finch genome with a single hit and an *E*-value $< 10^{-40}$ (Table S2, Supporting information).

### *Population genomics*

Overall, patterns of divergence were found to be highly heterogenous across the genome. Most of the highly diverged loci were located within distinct blocks or genomic islands with island size and frequency being greatest on the Z chromosome. The sliding window estimation of population genomic parameters revealed at least one island of divergence on 15 of the 23 chromosomes. A total of 5.7% of the autosomal genome was found within islands, and nine islands on the microchromosomes were located near what we infer to be the centromeres (chromosomes 10, 11, 14, 15, 17, 19, 20, 21 and 28; Fig. 2; Fig. S2, Supporting information) based upon comparisons with homologous chromosomes in the Zebra finch and *Ficedula* flycatchers (Ellegren *et al.* 2012). Autosomal island size ranged from 9.7 kb to 3.26 Mb, with an average size of 545 kb. In contrast, 19.7% of the Z chromosome was found within an island with island size ranging from 52.2 kb to 1.65 Mb with an average of 436 kb. A comparison of population genomic parameters within and outside islands revealed that $d_f$ and $F_{ST}$ were, respectively, >30 to >600 times higher within islands than outside islands on the autosomes, while $\pi'$ was ~two-fold lower within islands than outside and $d_{xy}$ was similar genome wide (Table 1). The Z chromosome showed elevated levels of differentiation in comparison with the autosomes with $F_{ST}$ values outside islands on the Z chromosome exceeding $F_{ST}$ outside islands on the autosomes by a factor of ~200, while $F_{ST}$ values within islands on the Z chromosome were ~four-fold higher than $F_{ST}$ values within islands on the autosomes (Table 1; Fig. 2).

### *Simulations to test for the influence of drift*

The frequency distribution of $F_{ST}$ values was strongly bi-modal with a first peak centred around $F_{ST} = 0$ and a second major peak at $F_{ST} = 1$ (Fig. 3). Coalescent simulations designed to assess the likelihood of observing the expected results via drift alone strongly support the idea that forces other than drift are driving divergence of loci with $F_{ST}$ >0.8 (Fig. 3). Clearly, any reasonable rate of neutral migration that matches the distribution

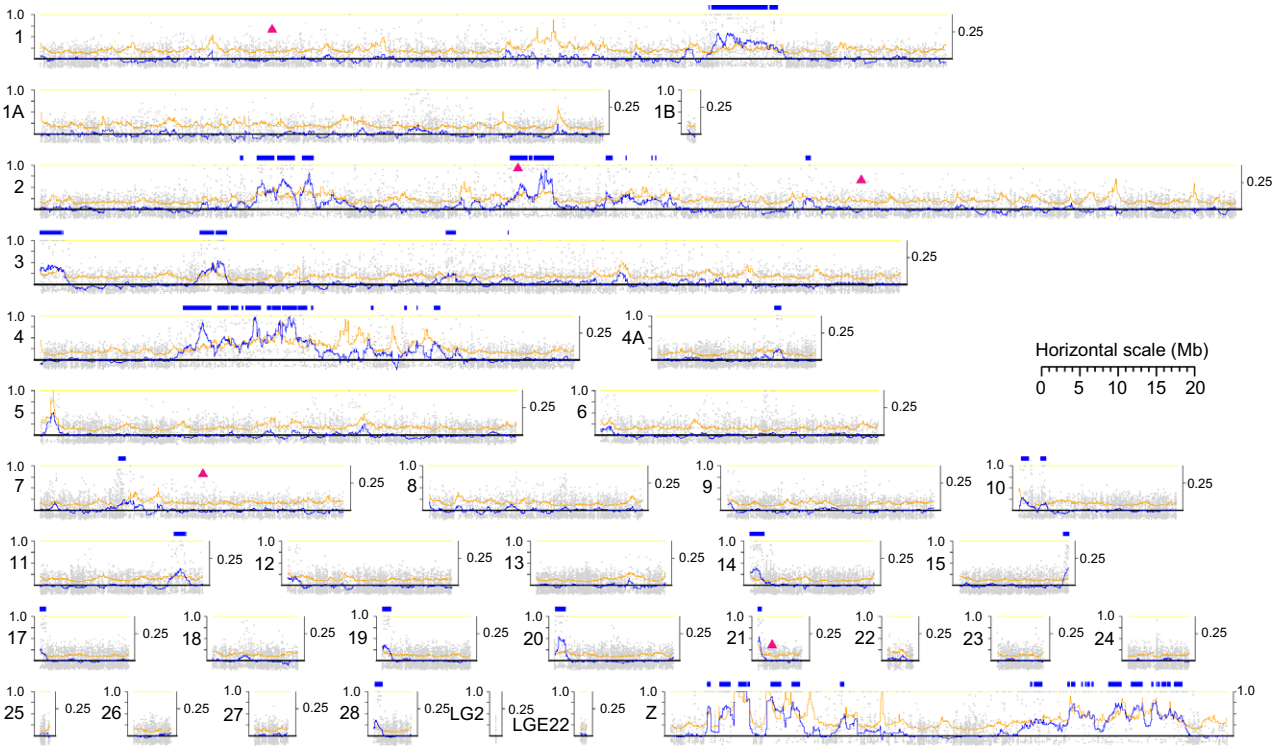


**Fig. 2** Genomic architecture of divergence in the Swainson's thrush. Grey dots represent the $F_{ST}$ values at each SNP across all 33 chromosomes. *X*-coordinates denote genomic locations. $F_{ST}$ values are scaled relative to the left-hand *y*-axis, with the horizontal yellow lines at $F_{ST} = 1$; blue line represents the smoothed $F_{ST}$ value using a window width of 1 Mb and its height is relative to the right-hand *y*-axis so that features are visible. The orange line represents the maximum smoothed value of $F_{ST}$ from 25 000 permutations simulated under the null hypothesis of no spatial autocorrelation (see Materials and methods). Solid blue horizontal bars denote regions with defined islands (observed, smoothed $F_{ST}$ exceeds all permuted, smoothed $F_{ST}$s). Red triangles denote the locations of the candidate migration genes (Table 2).

**Table 1** Mean values of population genomic parameters with standard errors inside and outside genomic islands

| Parameter | Autosomal background | SE | Autosomal island | SE | Z chromosome background | SE | Z chromosome island | SE |
|---|---|---|---|---|---|---|---|---|
| $d_f$ | 0.0022 | $2.9 \times 10^{-4}$ | 0.0785 | $7.1 \times 10^{-3}$ | 0.1335 | $1.2 \times 10^{-2}$ | 0.7962 | $5.9 \times 10^{-2}$ |
| $F_{ST}$ | 0.0002 | $4.2 \times 10^{-4}$ | 0.1293 | $4.7 \times 10^{-3}$ | 0.0374 | $4.1 \times 10^{-3}$ | 0.4946 | $3.1 \times 10^{-2}$ |
| $\pi_{ustulatus}$ | 0.0014 | $9.0 \times 10^{-6}$ | 0.0006 | $1.8 \times 10^{-5}$ | 0.0008 | $3.8 \times 10^{-5}$ | 0.0001 | $2.1 \times 10^{-5}$ |
| $\pi_{swainsoni}$ | 0.0018 | $9.4 \times 10^{-6}$ | 0.0007 | $2.1 \times 10^{-5}$ | 0.0014 | $4.7 \times 10^{-5}$ | 0.0003 | $3.5 \times 10^{-5}$ |
| $D_{xy}$ | 0.0018 | $9.7 \times 10^{-6}$ | 0.0015 | $4.0 \times 10^{-5}$ | 0.0017 | $5.4 \times 10^{-5}$ | 0.0026 | $1.8 \times 10^{-4}$ |

of $F_{ST}$ values <0.8 fails to produce the peaks of $F_{ST}$ >0.8 in the observed distribution. Furthermore, the simulation results make it clear that the levels of divergence found on the Z chromosome cannot be obtained merely by decreasing $N_e$ by a factor of 1/4.

### *Gene ontology and candidate gene analysis*

Overall, 13 543 contigs containing 60 783 SNPs mapped to 5183 known, novel or predicted genes within the Zebra finch genome. Of these, 670 genes were determined to be within an island of divergence, 131 genes were found to contain fixed differences between coastal and inland forms, and 65 of the genes with fixed differences were found within islands. An assessment of GO terms associated with genes in islands revealed no enrichment of any functional category after accounting for multiple comparisons ($P > 0.05$). An assessment of GO terms associated with SNPs with fixed differences between the coastal and inland forms revealed five functional categories with significant enrichment ($P < 0.05$; Table S3, Supporting information), but these were not known to be associated with traits for which we had an a priori reason to believe were involved in adaptive divergence.

To assess the potential that genes linked to traits known to be involved in reproductive isolation in birds, such as song, plumage coloration and migratory behaviour, may be more diverged than expected by chance, we conducted a literature review and identified 18 genes for which we had sequence data (Table S4, Supporting information; Table 2). Of these, only one gene linked to migration (CPNE4) fell within an island of divergence (Table 2; Fig. 3). However, the occurrence of high-$F_{ST}$ SNPs within the candidate migration genes was highly significantly more frequent than expected by chance alone ($P < 10^{-4}$; Table 2; Fig. 4). In contrast, the occurrence of high-$F_{ST}$ SNPs in the candidate song and colour genes was not significantly more frequent than expected by chance alone (song $P = 0.67$, colour $P = 0.50$; Fig. 3; Fig. S3, Supporting information). Because the song genes were selected from a large list of possible candidates, we further explored the possibility that our results could have been influenced by the subset of genes selected. To test this idea, we conducted an analysis of a larger set of genes ($n = 40$) identified by table 6 in Nam *et al.* (2010) and Table S7 (Supporting information) in Warren *et al.* (2010). While the larger subset of candidate song genes contained some individual outliers, the overall pattern of high-$F_{ST}$ SNPs within

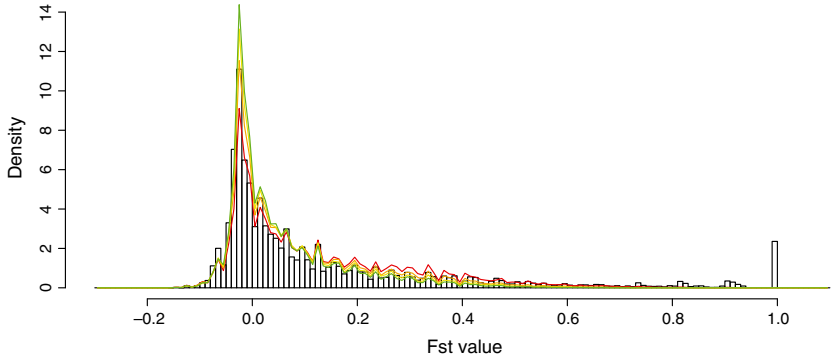


**Fig. 3** Coalescent simulations assessing the extent to which our results can be explained by drift. Green, yellow, orange and red lines represent the distribution of $F_{ST}$ values simulated under a neutral island model of migration with 4 $N_em$ = 2, 3, 4 and 5, respectively. The underlying histogram represents the distribution of observed $F_{ST}$ values. The comparison between the observed and expected values suggests that the large number of values >0.8 cannot be explained by drift alone.

**Table 2** Patterns of differentiation in genes linked to migration, song and colour in birds (see Table S4, Supporting information for additional information)

| Gene name | Hypothesized function | Max $F_{ST}$ | In island (Y/N) | Chromosome |
|---|---|---|---|---|
| Migration/circadian rhythm | | | | |
| CPNE4 | Migratory sleeplessness | 0.93 | Y | 2 |
| CREB1 | Circadian rhythm | 0.82 | N | 7 |
| NPAS2 | Circadian rhythm | 0.66 | N | 1 |
| ADCYAP1 | Migratory restlessness | 0.66 | N | 2 |
| PER3 | Circadian rhythm | 0.35 | N | 21 |
| Neurobiology of vocal communication | | | | |
| CACNA1B | Vocal communication | 0.4 | N | 17 |
| CACNA1G | Vocal communication | 0.26 | N | 18 |
| KCNC2 | Vocal communication | 0.22 | N | 1A |
| GRIA2 | Vocal communication | 0.27 | N | 4 |
| GRIA3 | Vocal communication | 0.27 | N | 4A |
| Pigmentation | | | | |
| ASIP | Dorsal ventral pigmentation | −0.11 | N | 20 |
| BCMO1 | Carotenoid breakdown | 0.1 | N | 11 |
| DCT/TYRP2 | Codes for melanogenesis enzymes | 0.25 | N | 1 |
| GSTA2/B5FYW2_TAEGU | Carotenoid binding | 0.09 | N | 3 |
| MLN64/STARD3 | Carotenoid deposition | 0.15 | N | 27 |
| SR-BI/SCARB1 | Carotenoid uptake | 0.24 | N | 15 |
| StAR4/STARD4 | Carotenoid deposition | 0.84 | N | Z |
| TYR | Codes for melanogenesis enzymes | 0.41 | N | 1 |

the song genes was not significantly different than expected by chance ($P = 0.30$).

## Discussion

The availability of next-generation sequencing data has raised new questions about the process of divergence at the level of the genome. One topic that has received considerable attention is the role of genomic islands during the process of speciation with ongoing gene flow (Turner *et al.* 2005; Via 2009; Michel *et al.* 2010; Feder *et al.* 2012a,b; Nosil & Feder 2012), versus when gene flow is absent or intermittent (Ellegren *et al.* 2012; Nadeau *et al.* 2012; Renaut *et al.* 2013). Here, we use a recently developed approach to create long contigs from short-read sequence data, RAD-PE sequencing (Etter *et al.* 2011), to investigate the genomic landscape of divergence between subspecies of a migratory bird experiencing limited gene flow along a putative zone of secondary contact. This study complements other recent work on the genomic landscape of divergence between groups with more complex evolutionary histories (Renaut *et al.* 2012, 2013; Stolting *et al.* 2013) and provides novel insights into the genes and gene regions involved in the early stages of divergence in a migratory songbird.

### *The origin of genomic islands of divergence*

We provide clear evidence for strong heterogeneity in patterns of genome-wide divergence between subspecies of the Swainson's thrush, with multiple, large blocks of physically linked outliers identified across the genome (average $\text{length}_{auto}$ ~545 kb, average $\text{length}_{Z}$ = 436 kb; Fig. 2). Our coalescent simulations indicate that a simple neutral model of genetic drift cannot explain the observed patterns of differentiation (Fig. 3), and lower nucleotide diversity within the islands is consistent with the idea that non-neutral processes play a role in their origin (Fig. S2, Supporting information). While the lack of a karyotype for the Swainson's thrush limits our ability to draw strong conclusions regarding the role of genomic architecture in the formation of genomic islands, we find a high concentration of genomic islands near what we infer to be the centromeres based upon comparisons with homologous chromosomes in *Ficedula* flycatchers and the Zebra finch (Ellegren *et al.* 2012). This pattern is particularly apparent on the microchromosomes that are generally thought to be telocentric in birds (chromosomes 10, 11, 14, 15, 17, 19, 20, 21 and 28) (Shields 1982). Centromeres are associated with suppressed recombination and complex repeat structures, both of which are thought to promote diver-

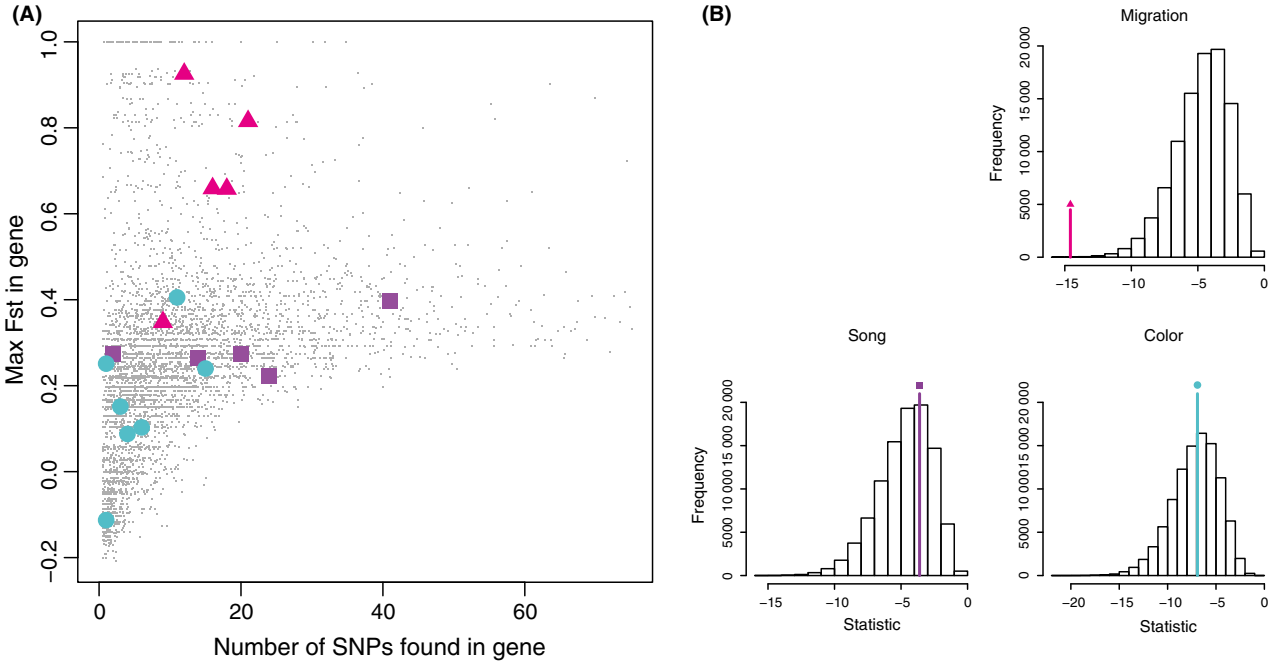


**Fig. 4** Comparative candidate gene analysis indicating that candidate migration genes are more diverged than candidate song and colour genes. (A) Red triangles are migration-related genes, blue circles are colour-related genes, and purple squares are song-related genes. Grey dots are all other SNPs genotyped in autosomal coding regions. (B) Test statistic indicating the likelihood of observing the patterns by chance alone. The chance of observing four out of five migration-related genes as outliers is <1 in 10 000. The observed patterns for song and colour do not deviate from random expectations. Because evolution was faster on the Z chromosome, the results are shown separately (Fig. S3, Supporting information).

gence (Carneiro *et al.* 2009; Ellegren *et al.* 2012). Overall, our results support the idea that differentiation between groups with complex evolutionary histories is highly heterogeneous and accelerated by both selection and aspects of genome architecture.

Speciation-with-gene-flow models predict that genomic island may arise as a result of selection at loci involved in reproductive isolation, while gene flow will break apart linkages in the surrounding regions (Noor *et al.* 2001; Navarro & Barton 2003). However, in the Swainson's thrush, two lines of evidence run counter to a speciation-with-gene-flow model for the formation of the observed genomic islands. First, if contemporary gene flow following secondary contact were responsible for the evolution of genomic islands in this species, we would expect greater homogeneity in patterns of divergence between populations located far from the hybrid zone. However, it is difficult to imagine a scenario in which patterns of divergence are more heterogeneous than those observed here (Fig. 2), despite the fact that the populations sampled in this study were from the extreme ends of the species range where the effect of contemporary gene flow is minimal (Fig. 1). Second, we found little support for the idea that genes influencing traits known to contribute to reproductive isolation in songbirds are concentrated within genomic islands. We found no significant enrichment of GO terms in genes found within the islands and all except for one of the candidate genes identified by our literature review fell outside the genomic islands. Thus, the results presented here support a limited role for current gene flow in explaining the origin of the observed genomic islands of divergence.

An alternative explanation for the origin of the genomic islands is that selection has occurred on the same regions in both lineages independently as a result of features inherent to the underlying genomic architecture (Amores *et al.* 2011; Renaut *et al.* 2013). Regions of low recombination, such as the centromeres, are more prone to selective sweeps and/or negative background selection at loci that may or may not be involved in reproductive isolation (Charlesworth *et al.* 1993; Charlesworth 1998; Noor & Bennett 2010). Furthermore, regions of low differentiation around islands may result from shared variation pre-dating the population split and/or introgression following previous episodes of secondary contact (Stolting *et al.* 2013). In the future, it will be important to compare our data (from geographically disparate populations) to patterns of divergence between populations located in close proximity to the hybrid zone. If the size and location of genomic islands are the same regardless of geography, it would further support the idea that regions of high diver-

gence may arise in the two separate lineages independently as a result of selection and features of the underlying genomic architecture. Furthermore, with increased sampling from across the range, it will also be possible to test the origin of low-divergence regions using statistical methods designed to distinguish between shared variation pre-dating the subspecies split and gene flow (Wakeley & Hey 1997; Joly *et al.* 2009; Sousa & Hey 2013; Seehausen *et al.* 2014).

### *Genes involved in divergence in a migratory songbird*

While we found no evidence that candidate genes associated with behaviours and traits believed to be of adaptive consequence in songbirds were concentrated within islands, we did find that genes linked to migratory behaviour were significantly more divergent than expected by chance ($P < 10^{-4}$; Fig. 4; Table 2) – a result that is in keeping with our current understanding of traits important to the ecology and evolutionary history of this species. The two subspecies of the Swainson's thrush are known to display strong differences in the length, timing and direction of migration (Ruegg & Smith 2002; Ruegg 2008; Delmore *et al.* 2012), and it has been hypothesized that differences in migratory behaviour may lead to some reproductive isolation across hybrid zones between distinct migratory forms in this and other species (Bearhop *et al.* 2005; Ruegg *et al.* 2012).

Not surprisingly, the candidate migration genes included in our study are linked to the circadian clock (Johnsen *et al.* 2007; Jones *et al.* 2008; Steinmeyer *et al.* 2009; Mueller *et al.* 2011), a biological process believed to control many components of migratory behaviour, including the timing, extent and duration of migration (Gwinner 1977, 1996; Bartell & Gwinner 2005). One of the genes showing significant divergence in our analysis, ADCYAP1, has also been found to be associated with migratory restlessness in blackcaps (Mueller *et al.* 2011). Additionally, the gene demonstrating the strongest pattern of divergence between *C. u. ustulatus* and *C. u. swainsoni*, CPNE4, was differentially expressed in migratory versus nonmigratory white crowned sparrows (Jones *et al.* 2008). The remaining two outliers in our study, CREB1 and NPAS2, have been linked to the circadian rhythms in other organisms, but no associations with migratory behaviour in birds have previously been identified (Steinmeyer *et al.* 2009). Overall, significant variation at genes linked to the circadian clock is in keeping with the idea that migration is involved in the early stages of divergence in migratory birds. Future work will assess patterns of divergence in these genes across the hybrid zone to inform our understanding of their potential role in maintaining the barrier to gene flow between the subspecies.

We also assessed differentiation in genes linked to song and plumage coloration, two other characters that are known to differ between the subspecies, but found that they were not more divergent than one would predict by chance alone (song $P = 0.67$, colour $P = 0.50$; Fig. 3). Lack of significant divergence in genes linked to song is perhaps not surprising given the potential for cultural transmission of acoustic characters in many songbirds (Slater 1986; Catchpole & Slater 2003). Some evidence suggests that characters such as song frequency are heritable in species with vocal learning, but this may be largely due to correlations with body size (Forstmeier *et al.* 2009), and the extent to which song variation is culturally inherited in the Swainson's thrush is not known. Similarly, the extent to which plumage coloration is influenced by genetic versus environmental factors is also unknown. For both traits, it is also possible that there is a genetic component to their inheritance, but that it is mediated by genes other than the ones included in the present study. Furthermore, all of our analyses were limited to the genes for which we had RAD-seq data and a more complete survey of all available candidate genes may yield a different result. In the future, experiments linking song behaviour and pigmentation with gene expression will improve our understanding of the genetic basis of vocalization and coloration in birds and the potential role of these traits in the process of divergence.

### *The particular role of the avian sex chromosomes*

It has long been hypothesized that sex chromosomes may play a disproportionately large role in speciation (Dobzhansky 1974; Coyne 1985; Charlesworth *et al.* 1987; Ellegren 2009). Alternatively, because the Z chromosome has 3/4 the effective population size of an autosome, it has also been suggested that genetic drift may increase the rate of evolution in this region, even in the absence of selection (Mank *et al.* 2010). Here, we find that divergence within and outside islands on the Z chromosome is highly elevated beyond what would be expected by drift alone (Fig. 3). While additional data are necessary to distinguish between alternative selection-based hypotheses for faster evolution of the sex chromosomes (Charlesworth *et al.* 1987; Iyengar *et al.* 2002; Begun *et al.* 2007; Mank *et al.* 2007), we note that some of the GO terms associated with genes containing fixed differences on the Z chromosome include functions such as embryo implantation, sexual reproduction, male courtship behaviour and penetration of the zona pellucida. Further experiments will reveal the extent to which fixed differences in traits related to sexual reproduction may play a role in the formation of reproductive barriers between *C. u. ustulatus* and *C. u. swainsoni*.

*Individual-level RAD-PE sequencing for population genomics*

In addition to assessing the genomic landscape of divergence in a migratory bird, we also demonstrate the utility of RAD-PE sequencing and mapping to the Zebra finch genome for SNP discovery, quality control, and gene ontology analysis. RAD-PE sequencing and mapping to the Zebra finch genome using a series of stringent quality filters (see Materials and methods) resulted in 33 841 contigs (average ~300 bp) containing 154 123 SNPs, 13 543 of which could be mapped to known or predicted genes in the Zebra finch genome (Table S2, Supporting information). This data set represents, to the best of our knowledge, the largest avian SNP data set generated using RAD sequencing. Despite the fact that rearrangements and inversions since the time of divergence with the Zebra finch may confound the true location of some of our SNP loci, we were still able to detect strong spatial autocorrelation in patterns of divergence. While estimates of divergence times between the Zebra finch and Swainson's thrush range from 39.2 to 72 mya (Brown *et al.* 2008), it is well-accepted that synteny is highly conserved in Passeriformes (Derjusheva *et al.* 2004; Griffin *et al.* 2008) and our data add to the growing number of studies demonstrating the utility of the Zebra finch genome as a rough backbone for mapping in the absence of a species-specific songbird genome (Ellegren *et al.* 2012; Bourgeois *et al.* 2013). However, unlike previous studies that used RAD-PE sequencing on pooled samples (Bourgeois *et al.* 2013), here, we generate RAD-PE data for individually barcoded samples. Individual barcoding makes it possible to avoid potential biases associated with sample pooling, including over- or under-representation of individuals in a pool due to variation in sample concentration, sample quality or random processes (Anderson *et al.* 2014) and allows us to assess allele frequencies across the genome more reliably. As sequencing technology improves, use of RAD-PE sequencing on individually barcoded samples will be a good choice for population genomic studies requiring accurate and robust allele frequency estimates at individual loci.

## Conclusions

Here, we investigate the genomic landscape of divergence in a migratory songbird, the Swainson's thrush, to test emerging theories about the process of speciation at the level of the genome. Our analyses reveal that loci linked to migratory behaviour were significantly more divergent than expected by chance—a particularly salient finding in the light of the documented differences in migratory behaviour between the subspecies and the potential role of migration in limiting gene flow across the hybrid zone. Counter to the expectations of a divergence-with-gene-flow model, the most diverged migration candidate genes were not located within prominent islands of genomic divergence, indicating that divergence in migration and the formation of genomic islands (possibly due to selective sweeps, background selection and differences in genomic architecture) may arise from entirely separate processes. Evidence for fast evolution on the Z chromosome—far above what would be expected by drift alone—further solidifies the idea that sex chromosomes play a large role in the early stages of divergence in birds. Furthermore, we find that RAD-PE sequencing of individually barcoded samples and alignment to the Zebra finch genome provides a good option for population genomic studies requiring accurate and robust allele frequency estimates at individual loci, but where genomic resources may be limited. Overall, our results lend valuable insight into the multifaceted mechanisms underlying speciation at the level of the genome and provide new data to help explain avian species diversity.

## Acknowledgements

This research was supported by a donation from Margery Nicolson and a grant to TBS from The Turner Foundation, EPA (RD-83377801). We would like to thank Craig Moritz and Daniel Barshis for helpful discussions on earlier versions of this work and Tressa Atwood and Rick Nipper for their assistance with the laboratory and bioinformatics components. We would also like to thank Chris Witt at the Museum of Southwestern Biology and Peter Marra of the Smithsonian Migratory Bird Center for contributing genetic material. The authors would like to thank the thorough and helpful suggestions provided by the anonymous reviewers.

## References

Amores A, Catchen J, Ferrara A, Fontenot Q, Postlethwait JH (2011) Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, **188**, 799–808.

Anderson EC, Skaug HJ, Barshis DJ (2014) Next-generation sequencing for molecular ecology: a caveat regarding pooled samples. *Molecular Ecology*, **23**, 502–512.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One*, **3**, e3376.

Bartell PA, Gwinner E (2005) A separate circadian oscillator controls nocturnal migratory restlessness in the songbird Sylvia borin. *Journal of Biological Rhythms*, **20**, 538–549.

Bearhop S, Fiedler W, Furness RW *et al.* (2005) Assortative mating as a mechanism for rapid evolution of a migratory divide. *Science*, **310**, 502–504.

Begun DJ, Holloway AK, Stevens K *et al.* (2007) Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biology*, **5**, 2534–2559.

Bourgeois YXC, Lhuillier E, Cézard T *et al.* (2013) Mass production of SNP markers in a nonmodel passerine bird through RAD sequencing and contig mapping to the zebra finch genome. *Molecular Ecology Resources*, **13**, 899–907.

Brown JW, Rest JS, Garcia-Moreno J, Sorenson MD, Mindell DP (2008) Strong mitochondrial DNA support for a Cretaceous origin of modern avian lineages. *BMC Biology*, **6**, 6.

Carneiro M, Ferrand N, Nachman MW (2009) Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (*Oryctolagus cuniculus*). *Genetics*, **181**, 593–606.

Catchpole CK, Slater PJB (2003) *Bird Song: Biological Themes and Variations*. Cambridge University Press, New York.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.

Charlesworth B, Coyne JA, Barton NH (1987) The relative rates of evolution of sex chromosomes and autosomes. *The American Naturalist*, **130**, 113–146.

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

Coyne JA (1985) The genetic basis of Haldanes rule. *Nature (London)*, **314**, 736–738.

Coyne JA, Orr AH (2004) *Speciation*. Sinauer Associates Inc, Sunderland.

Delmore KE, Fox JW, Irwin DE (2012) Dramatic intraspecific differences in migratory routes, stopover sites and wintering areas, revealed using light-level geolocators. *Proceedings, Biological Sciences*, **279**, 4582–4589.

Derjusheva S, Kurganova A, Habermann F, Gaginskaya E (2004) High chromosome conservation detected by comparative chromosome painting in chicken, pigeon and passerine birds. *Chromosome Research*, **12**, 715–723.

Dobzhansky T (1974) Genetic analysis of hybrid sterility within the species Drosophila–Pseudoobscura. *Hereditas*, **77**, 81–88.

Ellegren H (2009) Genomic evidence for a large-Z effect. *Proceedings, Biological Sciences*, **276**, 361–366.

Ellegren H, Smeds L, Burri R *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 4.

Etter PD, Preston JL, Bassham S, Cresko WA, Johnson EA (2011) Local de novo assembly of RAD paired-end contigs using short sequencing reads. *PLoS One*, **6**, e18561.

Evans Mack D, Yong W (2000) Swainson's thrush (*Catharus ustulatus*). In: *The Birds of North America* (eds Poole A & Gills F), pp. 1–19. Birds of North America Inc., Philadelphia, PA.

Feder JL, Egan SP, Nosil P (2012a) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342–350.

Feder JL, Gejji R, Yeaman S, Nosil P (2012b) Establishment of new mutations under divergence and genome hitchhiking. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 461–474.

Forstmeier W, Burger C, Temnow K, Deregnaucourt S (2009) The genetic basis of zebra finch vocalizations. *Evolution*, **63**, 2114–2130.

Griffin DK, Robertson LB, Tempest HG *et al.* (2008) Whole genome comparative studies between chicken and turkey and their implications for avian genome evolution. *BMC Genomics*, **9**, 168.

Gwinner E (1977) Circannual rhythms in bird migration. *Annual Review of Ecology and Systematics*, **8**, 381–405.

Gwinner E (1996) Circannaual clocks in avian reproduction and migration. *Ibis*, **138**, 47–63.

Hohenlohe PA, Bassham S, Etter PD *et al.* (2010) Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Iyengar VK, Reeve HK, Eisner T (2002) Paternal inheritance of a female moth's mating preference. *Nature (London)*, **419**, 830–832.

Johnsen A, Fidler AE, Kuhn S *et al.* (2007) Avian Clock gene polymorphism: evidence for a latitudinal cline in allele frequencies. *Molecular Ecology*, **16**, 4867–4880.

Joly S, McLenachan PA, Lockhart PJ (2009) A statistical approach for distinguishing hybridization and incomplete lineage sorting. *The American Naturalist*, **174**, E54–E70.

Jones S, Pfister-Genskow M, Cirelli C, Benca RM (2008) Changes in brain gene expression during migration in the white-crowned sparrow. *Brain Research Bulletin*, **76**, 536–544.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Li H, Handsaker B, Wysoker A *et al.* (2009) The sequence alignment/map format and SAMTOOLS. *Bioinformatics*, **25**, 2078–2079.

Mallet J, Beltran M, Neukirchen W, Linares M (2007) Natural hybridization in heliconiine butterflies: the species boundary as a continuum. *BMC Evolutionary Biology*, **7**, 28.

Mank JE, Axelsson E, Ellegren H (2007) Fast-X on the Z: rapid evolution of sex-linked genes in birds. *Genome Research*, **17**, 618–624.

Mank JE, Nam K, Ellegren H (2010) Faster-Z evolution is predominantly due to genetic drift. *Molecular Biology and Evolution*, **27**, 661–670.

Metzker ML (2010) Sequencing technologies—the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Michel AP, Sim S, Powell THQ *et al.* (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 9724–9729.

Mueller JC, Pulido F, Kempenaers B (2011) Identification of a gene associated with avian migratory behaviour. *Proceedings, Biological Sciences*, **278**, 2848–2856.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing *Heliconius butterflies* identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 343–353.

Nam K, Mugal C, Nabholz B *et al.* (2010) Molecular evolution of genes in avian genomes. *Genome Biology*, **11**, R68.

Navarro A, Barton NH (2003) Accumulating postzygotic isolation genes in parapatry: a new twist on chromosomal speciation. *Evolution*, **57**, 447–459.

Noor MAF, Bennett SM (2010) Islands of speciation or mirages in the desert: examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

Noor MA, Cunningham AL, Larkin JC (2001) Consequences of recombination rate variation on quantitative trait locus mapping studies. Simulations based on the *Drosophila melanogaster* genome. *Genetics*, **159**, 581–588.

Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. *Philosophical Transactions of the Royal Society of London, Series B Biological Sciences*, **367**, 332–342.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375–402.

Price T (2008) *Speciation in Birds*. Roberts and Company Publishers, Colorado.

Renaut S, Maillet N, Normandeau E *et al.* (2012) Genome-wide patterns of divergence during speciation: the lake whitefish case study. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences*, **367**, 354–363.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Ruegg K (2008) Genetic, morphological, and ecological characterization of a hybrid zone that spans a migratory divide. *Evolution*, **62**, 452–466.

Ruegg KC, Smith TB (2002) Not as the crow flies: a historical explanation for circuitous migration in Swainson's thrush (*Catharus ustulatus*). *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **269**, 1375–1381.

Ruegg KC, Hijmans RJ, Moritz C (2006a) Climate change and the origin of migratory pathways in the Swainson's thrush (*Catharus ustulatus*). *Journal of Biogeography*, **33**, 1172–1182.

Ruegg KC, Slabbekoorn H, Clegg SM, Smith TB (2006b) Divergence in mating signals correlates with ecological variation in a migratory songbird, the Swainson's thrush (*Catharus ustulatus*). *Molecular Ecology*, **15**, 3147–3156.

Ruegg K, Anderson EC, Slabbekoorn H (2012) Differences in timing of migration and response to sexual signalling drive asymmetric hybridization across a migratory divide. *Journal of Evolutionary Biology*, **25**, 1741–1750.

Seehausen O, Butlin RK, Keller I *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176–192.

Shields GF (1982) Comparative avian cytogenetics: a review. *Condor*, **84**, 45–58.

Slater PJ (1986) The cultural transmission of bird song. *Trends in Ecology & Evolution*, **1**, 94–97.

Smadja CM, Canback B, Vitalis R *et al.* (2012) Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. *Evolution*, **66**, 2723–2738.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404–414.

Steinmeyer C, Mueller JC, Kempenaers B (2009) Search for informative polymorphisms in candidate genes: clock genes and circadian behaviour in blue tits. *Genetica*, **136**, 109–117.

Stolting KN, Nipper R, Lindtke D *et al.* (2013) Genomic scan for single nucleotide polymorphisms reveals patterns of divergence and gene flow between ecologically divergent species. *Molecular Ecology*, **22**, 842–855.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biology*, **3**, 1572–1578.

Via S (2009) Natural selection in action during speciation. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 9939–9946.

Wakeley J, Hey J (1997) Estimating ancestral population parameters. *Genetics*, **145**, 847–855.

Warren WC, Clayton DF, Ellegren H *et al.* (2010) The genome of a songbird. *Nature*, **464**, 757–762.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

Wu X, Watson M (2009) CORNA: testing gene lists for regulation by microRNAs. *Bioinformatics*, **25**, 832–833.

Yeaman S (2013) Genomic rearrangements and the evolution of clusters of locally adaptive loci. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, E1743–E1751.

Zerbino DR, Birney E (2008) VELVET: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, **18**, 821–829.

K.R. and T.S. conceived of the project. K.R. oversaw the analysis, and wrote the paper. E.A. and J.P. wrote the scripts for the population genomic analysis and created the figures. J.B. oversaw the RAD sequencing and de novo assembly of paired-end contigs.

## Data accessibility

The data and scripts for population genomic analysis (from SNP selection to estimation of population genomic parameters) are available on Dryad (doi:10.5061/dryad.73gj4).

## Supporting information

Additional supporting information may be found in the online version of this article.

**Fig. S1** Length of paired-end aligned contigs.

**Fig. S2** Distribution of population genomic parameters, including: (i) the density of fixed differences per base pair ($d_f$); (ii) the between-species population differentiation ($F_{ST}$); (iii) the zero-truncated within-species nucleotide diversity (the average number of pairwise differences between contigs that carry at least one SNP), $\pi'$, and; (iv) the between-species average number of pairwise differences ($d_{xy}$).

**Fig. S3** Comparative candidate gene analysis similar to Fig. 3, but including the SNPs on the Z chromosome.

**Table S1** RAD sequencing results.

**Table S2** Counts of RAD-PE contigs and putative SNPs after each filtering step.

**Table S3** Go term analysis of genes containing max $F_{ST}$ SNPs = 1 between coastal and inland groups.

**Table S4** Candidate genes for migration, song, and color based upon a literature review.

**Appendix S1** Supplementary methods.