PROCEEDINGS B

rspb.royalsocietypublishing.org

Research

Cite this article: Bay RA, Ruegg K. 2017 Genomic islands of divergence or opportunities for introgression? *Proc. R. Soc. B* **284**: 20162414.
http://dx.doi.org/10.1098/rspb.2016.2414

Received: 3 November 2016
Accepted: 10 February 2017

Subject Category:
Evolution

Subject Areas:
genomics, evolution

Keywords:
speciation, genomic islands of divergence, gene flow, introgression

Author for correspondence:
Rachael A. Bay
e-mail: rachaelbay@gmail.com

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3700144.

THE ROYAL SOCIETY
PUBLISHING

Genomic islands of divergence or opportunities for introgression?

Rachael A. Bay and Kristen Ruegg

Center for Tropical Research, Institute for the Environment and Sustainability, University of California Los Angeles, Los Angeles, CA, USA

RAB, 0000-0002-9516-5881

In animals, introgression between species is often perceived as the breakdown of reproductive isolating mechanisms, but gene flow between incipient species can also represent a source for potentially beneficial alleles. Recently, genome-wide datasets have revealed clusters of differentiated loci ('genomic islands of divergence') that are thought to play a role in reproductive isolation and therefore have reduced gene flow. We use simulations to further examine the evolutionary forces that shape and maintain genomic islands of divergence between two subspecies of the migratory songbird, Swainson's thrush (*Catharus ustulatus*), which have come into secondary contact since the last glacial maximum. We find that, contrary to expectation, gene flow is high within islands and is highly asymmetric. In addition, patterns of nucleotide diversity at highly differentiated loci suggest selection was more frequent in a single ecotype. We propose a mechanism whereby beneficial alleles spread via selective sweeps following a post-glacial demographic expansion in one subspecies and move preferentially across the hybrid zone. We find no evidence that genomic islands are the result of divergent selection or reproductive isolation, rather our results suggest that differentiated loci both within and outside islands could provide opportunities for adaptive introgression across porous species boundaries.

1. Introduction

Disentangling conditions that lead to the formation and maintenance of new species has long been a central focus of evolutionary biology. Mayr [1] defined species as 'interbreeding natural populations that are reproductively isolated from other such groups', and since that time, introgression is often portrayed as a mechanism causing the breakdown of reproductive isolation. This idea has in turn shaped our interpretation of genomic differences between species. 'Islands of divergence' or clusters of highly differentiated loci have been described in numerous taxa [2–4], although patterns are not consistent across all systems [5]. In many cases, these regions are thought to have arisen at loci under divergent selection or involved in reproductive isolation, whereas gene flow works to homogenize the rest of the genome [2,3,6]. This 'divergence with gene flow' model has often been invoked in cases of ecological speciation [3,6–8]. For example, the most differentiated regions between *Heliconius* butterflies contain genes for colour pattern, which is under strong divergent selection and is involved in mate choice [4].

An alternative hypothesis for the formation of genomic islands of divergence is that they arose via selection in allopatry. Here, selection on a few beneficial alleles occurs in one or both populations and ultimately leads to genomic regions of differentiation between groups, accelerating the evolution of genomic islands [6,9,10]. In both selection in allopatry and divergence with gene flow models, high levels of linkage disequilibrium promote hitchhiking and therefore broaden regions of differentiation. Furthermore, aspects of genomic architecture such as proximity to centromeres can contribute to reduced recombination rates, resulting in slower breakdown of linkage disequilibrium between loci within islands and higher susceptibility to selective sweeps and/or background selection [5,10–12].

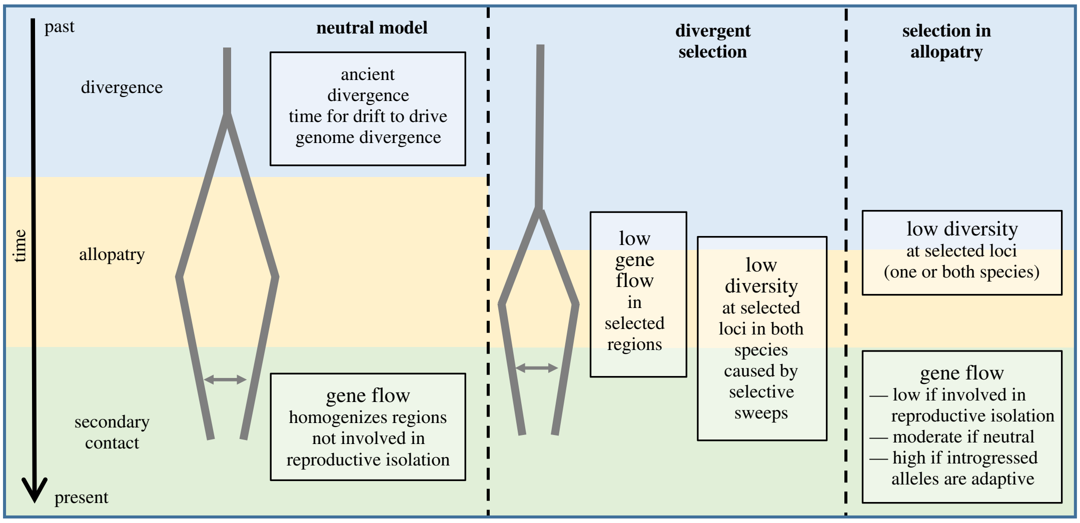
past
time
present
divergence
allopatry
secondary contact
neutral model
ancient divergence time for drift to drive genome divergence
gene flow homogenizes regions not involved in reproductive isolation
divergent selection
low gene flow in selected regions
low diversity at selected loci in both species caused by selective sweeps
selection in allopatry
low diversity at selected loci (one or both species)
gene flow
— low if involved in reproductive isolation
— moderate if neutral
— high if introgressed alleles are adaptive

Figure 1. Models explain origin and maintenance of genomic islands of divergence. (Online version in colour.)

Regions of secondary contact offer an opportunity for loci to be tested on new genetic backgrounds and in new environments, and traits that introgress between population or species boundaries can be adaptive. For example, levels of introgression between Darwin's finch species with different sized beaks are thought to be highly adaptive during severe droughts. In this case, beak size is so highly selected during drought seasons that hybrids are favoured and gene flow levels between species greatly increase [13,14]. In species with genomic islands of divergence, introgression and subsequent recombination will test these differentiated loci in new genomic and environmental contexts. If islands represent opportunities for adaptive introgression, then counter to divergence with gene flow models, one would predict these divergent regions to exhibit higher than neutral levels of gene flow between species in the early stages of the speciation process. It is important to note, however, that this pattern would only be apparent very early in speciation as adaptive alleles are likely to sweep to fixation once gene flow commences.

Islands of divergence have been documented in a number of young species pairs or subspecies [15–17]. In these cases, some regions of the genome are quite differentiated despite moderate to high rates of gene flow across the majority of the genome. For example, regions of elevated differentiation are especially pronounced in young species of *Heliconius* butterflies and these regions often correspond to known colour markers, which are involved in mate recognition [4]. Across a broad range of taxa, highly differentiated regions seem to consistently be located close to centromeres as well as across the sex chromosomes, regions known to have lower recombination rates [2,15]. In addition to the traditionally measured F_{ST}, recent studies have used other measures of differentiation and divergence (e.g. d_{xy} or Fay and Wu's H) to begin elucidating the origins of genomic islands of divergence [5,12]. Still, it remains difficult to disentangle the likely complex scenarios of demography and selection underlying the formation and maintenance of these islands of divergence with genome scans alone.

Here we take advantage of the opportunity to examine patterns of genomic differentiation in a well-studied hybrid zone between subspecies that have diverged across large portions of their genome, but gene flow between the subspecies is still ongoing. Two known subspecies of Swainson's thrush (*Catharus ustulatus*) overlap on their breeding grounds in the Coastal Mountains of British Columbia [18]. Previous work suggests that the two subspecies, a coastal subspecies (*C. ustulatus ustulatus*) and an inland subspecies (*C. ustulatus swainsonii*), were isolated during the last glacial maximum and since that time, the inland population has expanded westward and the two have come into secondary contact, with 40% of birds caught in the centre of the hybrid zone being identified as hybrids [19,20]. The subspecies differ in a number of important ecological traits, including wintering location, migration timing, plumage colour and song [18,20,21]. A genomic scan between pure representatives of these subspecies showed clustering of differentiated (high F_{ST}) loci in 132 different regions across 16 chromosomes, with particularly high F_{ST} in migration-linked genes found both within and outside of islands [17]. This general pattern was later corroborated with whole genome re-sequencing, but in contrast to Ruegg *et al.* [17], Delmore *et al.* [22] found that putative migration-linked genes were concentrated within islands—a result that was likely owing to exact parameters used to define genomic islands in each study. While both studies hypothesized that the observed islands formed as a result of selective sweeps in allopatry, neither modelled demography, which would allow estimates of divergence times and the ability to reject alternative explanations.

Here we use demographic simulations to examine the origin of genomic islands of divergence in *C. ustulatus* and their possible role in reproductive isolation between incipient species. Using estimates of divergence time, effective population size and gene flow, we explicitly test the following predictions (figure 1): (i) if genomic islands are a result of neutral divergence in isolation and low divergence in interisland regions are the result of introgression following secondary contact, we would predict ancient divergence time estimates, allowing time for drift to drive divergence, and high levels of gene flow in non-island regions. (ii) If genomic islands arose owing to divergent selection and play a prominent role in reproductive isolation, we would expect little to no gene flow between subspecies within islands and low nucleotide diversity within islands as a result of selective sweeps in both groups. (iii) If islands arose as a consequence of selective sweeps in one or both subspecies commensurate with a rapid post-Pleistocene expansion, we would not necessarily expect depressed levels of gene flow within islands, and may expect higher levels of gene flow at loci that confer an adaptive advantage in both groups until differentiation is erased, and/or decreased levels of gene flow in loci important to reproductive isolation. Our combination of demographic analysis and forward simulation goes a step beyond traditional genome scans, allowing us to disentangle potential mechanisms shaping genomic divergence.

2. Methods

(a) Demographic analysis using RAD-Seq data

To investigate demographic processes related to divergence, we used a previously published RAD-Seq dataset consisting of 25 individuals (15 from the inland subspecies and 10 from the coastal subspecies) from five different locations across the species range [17]. Although individuals were caught at different locations across the breeding range, preliminary within-subspecies PCA and Hardy–Weinberg (data not shown) did not show population structure, so we treated each subspecies as a single 'population' for simulations. These data were used to analyse genome-wide patterns of differentiation (F_{ST}) and define 'island' regions where F_{ST} outliers clustered in non-random patterns. Overall, 574 257 SNPs were identified from 132 645 contigs. We used a quality-filtered subset, including 360 632 SNPs from 64 513 contigs [17]. We used per-site nucleotide diversity (π) and F_{ST} as calculated in Ruegg *et al*. [17] to test the relationship between differentiation and diversity. Comparing smoothed differentiation statistics to a permuted null distribution, Ruegg *et al*. [17] identified 132 islands of divergence, broadly distributed across the genome. We created four sets of SNPs for our analysis, (i) all SNPs (ii) island SNPs on autosomes, (iii) non-island SNPs on autosomes and (iv) Z chromosome SNPs. We analysed the Z chromosome separately from the autosomes because it is expected to have different evolutionary history (with three-fourths the effective population size) and it was found to be more differentiated between the two subspecies; 33 of the 132 islands of divergence were located on the Z chromosome [17]. Also note that the 'all SNPs' category contains SNPs not included in the other three categories, because chromosomal location and thus island status could only be identified for *de novo* contigs that mapped to the zebra finch genome.

We inferred demographic history from RAD-Seq SNPs using the joint allele frequency spectrum, implemented in δaδi [23]. We downsampled data to 15 chromosomes and used a folded allele frequency spectra, as we do not have appropriate information for determining the true ancestral allele. First, we tested a divergence with gene flow model against a null hypothesis of divergence without gene flow. For each of four SNP sets, we fit both models and estimated log likelihood scores. We used a likelihood ratio test (d.f. = 2) to determine whether the migration model was a better fit to the data.

For estimation of divergence time, population sizes and migration rates, we subsampled without replacement each SNP set 100 times to generate a distribution of parameter estimates. We sampled the data in this manner for two reasons: first, the built-in bootstrapping framework in δaδi assumes markers are not linked and while our data almost certainly violate this assumption we have little knowledge of the actual linkage structure. Second, because the four SNP sets have drastically different sizes, it is impossible to compare parameter estimates among them. For this reason, each subsampled dataset consisted of the same number of SNPs (2000 SNPs). Note that the demographic analysis on selected subsets of the genome directly violates the assumptions of demographic inference models and therefore will provide skewed estimates of demographic parameters. We therefore rely on analysis of all SNPs to provide the best estimates of necessary parameters, but our primary goal is to examine relative estimates of gene flow among genomic regions rather than come up with an actual estimate of gene flow for the whole genome.

First, we subsampled all SNPs and estimated theta ($4N_{\text{ref}}\mu$l) and time of divergence (T). For all parameter estimations, we assumed a mutation rate (μ) of 3.44×10^{-9}, based on that derived for the closest relative in [24], the medium ground finch, a generation time of 2 years [18], and a sequence length (L) of 135 903 bp, based on a per base pair SNP rate calculated by dividing the total length of the de novo RAD assembly by the total number of SNPs identified. Because divergence time and migration rate are conflated, we fixed theta and time of divergence at the median value estimated from the distribution generated by subsampling all SNPs. This allowed us to estimate population sizes and migration rates for each SNP category while assuming that divergence time is the same for all SNPs. We estimated these parameters for each of 100 subsamples for each of the four SNP categories. In figures, we show the median and 95% CIs from these estimated distributions. All scripts for analyses are available at http://github.com/rachaelbay/SWTH-demography.

(b) Sensitivity to subsampling methods

To ensure that our sampling methods did not cause artefacts owing to either biological or statistical linkage between SNPs, we reran all simulations with two further sampling regimes. For both regimes, we allowed only one SNP per RAD contig in an attempt to reduce bias from highly linked loci. First, we resampled 1000 SNPs, close to the maximum number that could be sampled without replacement. We also resampled 5000 SNPs with replacement to test whether subsample size would impact our conclusions. For parameter estimates, total length (L) was adjusted based on the level of resampling.

(c) Effective population size trajectory

Trajectories of effective population size reflect true expansions and declines, but also shifts in population structure. We therefore expect to observe a decline in effective population size at the time of subspecies divergence, giving us an additional estimate of the approximate timing of the subspecies split. We used pairwise sequential Markovian coalescent (PSMC) [25] analysis to estimate fluctuations in population size over time. For this, we used previously published whole genome sequencing data from a single individual within the coastal subspecies [22]. Illumina short reads from one fragment library and two separate mate pair libraries were downloaded from NCBI's sequence read archive (SRR1812116, SRR1812117, SRR1812165) and aligned to the reference genome assembly, downloaded from Dryad [22] using stampy [26] and created a consensus sequence with samtools [27]. We also filtered sites with lower than 10× coverage, as suggested in [24] for this type of analysis. For PSMC analysis [25], we used parameters identified by Nadachowska-Brzyska *et al*. [24] to be ideal for analysis of avian genomes (N30 −t5 −r5 −p 4 + 30 × 2 + 4 + 6 + 10) running 100 bootstrap replicates. We used a mutation rate of 3.44×10^{-9}, based on that derived for the closest relative in [24], the medium ground finch. A generation time of 2 years [18] along with the mutation rate were used for scaling of effective population size and timing.

(d) Forward simulations

We used forward simulations to further test the impacts of selection and gene flow on patterns of genomic differentiation. All simulations were executed in the program SLiM [28]. We began with a single population based on the effective population

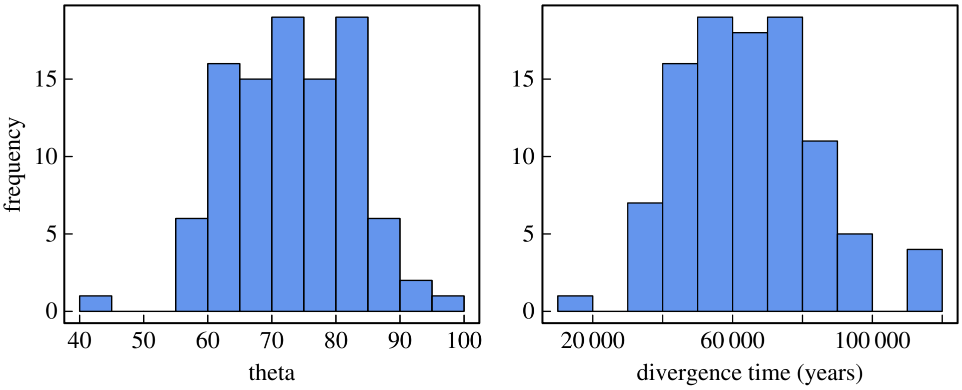

Figure 2. Distribution of parameter estimates for theta ($4N_{\mathrm{ref}}\mu l$) and divergence time estimated by subsampling all SNPs. Median values for these parameters are fixed in downstream modelling of gene flow across different SNP subsets. (Online version in colour.)

Table 1. Tests of divergence with gene flow model for four sets of SNPs: (i) all SNPs, (ii) autosomal SNPs in islands of divergence identified by Ruegg *et al.* [17], (iii) autosomal SNPs not in islands of divergence and (iv) SNPs located on the Z chromosome. A likelihood ratio test (d.f. = 2) was used to test whether a divergence with gene flow model is a better fit to the data than the null model without gene flow.

SNP subset	no SNPs	log likelihood (no migration)	log likelihood (migration)	*p*-value (LRT)
all	281 312	−27 416.8	−27 320.8	1.38×10^{-21}
island	5064	−950.3	681.6	4.56×10^{-59}
non-island	134 477	−10 041	−4279.3	0
Z-chromosome	3929	−939.2	−801.7	1.38×10^{-30}

size estimates (N_{ref}) from δaδi simulations, which underwent neutral mutation for 1000 generations to generate standing genetic variation. This population was then split into the two subspecies. Population expansion began after the last glacial maximum (18 000 years ago), and growth rates were calculated based on our estimates of current population size for each group. We simulated a single short chromosome (1 Mb) with recombination 1×10^{-6} and two 100 kb islands with a lower recombination rate of 1×10^{-9}; both rates are well within the normal range of recombination rates found in zebra finch [29]. These recombination rates were varied to test competing hypotheses about the role of recombination in generating clusters of highly differentiated loci. The mutation rate was 3.44×10^{-9} based on medium ground finch [24]. To decrease computation time involved in simulating large populations, we scaled all populations sizes, mutation rates, times, selective coefficients, and population growth rates by a factor of 10. We simulated three different selection scenarios: (i) neutral mutations only, (ii) selective sweeps in both subspecies and (iii) selective sweeps in inland subspecies only.

Selective sweeps were simulated by introducing one beneficial mutation each generation, with selection coefficient drawn from an exponential distribution with mean 0.05. We simulated timing of secondary contact under the most probable estimate of secondary contact, 6000 years ago, based upon palaeoecological-based estimates of the timing of suitable habitat in the region of the hybrid zone [20]. As an upper bound, we also simulated a very recent secondary contact, 100 years ago, based upon the first record of a hybrid from this hybrid zone [20]. We also simulated multiple migration rates, which were based on relative estimates from δaδi simulations (see electronic supplementary material for all parameter combinations). Finally, when selective sweeps were simulated in the inland subspecies only, we simulated scenarios where those same mutations were either beneficial or selectively neutral within the coastal subspecies. For each scenario, we calculated the F_{ST} and nucleotide diversity (calculated as expected heterozygosity—H_{E}) for each variable site.

3. Results

(a) Divergence and expansion

For all sets of SNPs, divergence with gene flow was a better fit to the data than the no gene flow model (LRT $p < 0.001$; table 1). Joint site frequency spectra are shown in electronic supplementary material, figure S1. Median divergence time based on 100 subsamples of 2000 SNPs was 64 610 years ago (95% CI = 34 595–112 224; figure 2). Estimates of divergence time based on subsamples of 1000 or 5000 SNPs were only slightly lower: 48 821 (95% CI = 17 486–318 018) and 48 864 (35 057–68 600) years, respectively. The historical trajectory of effective population size based on PSMC analysis of whole genome data from a single individual from the coastal subspecies shows a sharp decline beginning at slightly over 100 000 years ago (electronic supplementary material, figure S2). We hypothesize that this decline does not in fact correspond to a decrease in population size, but rather the subspecies split, though we recognize that PSMC results are likely confounded by ongoing gene flow. Although this estimate of divergence time is larger than that estimated from RAD-Seq data, all estimates are well within the Late Pleistocene and prior to the last glacial maximum. Median estimates of theta (defined in δaδi as $4N_{\mathrm{ref}}\mu l$) were 34.5 (95% CI = 12.43–24.29), 73.33 (56.83–90.42) and 168.5 (141.10–195.00) for the 1000, 2000 and 5000 SNP subsets. Median estimates of theta and divergence time for each sampling regime were fixed in further simulations with subsets of SNPs representing different genomic regions.

Estimates of effective population size show evidence of population expansion in both groups. Based on our estimate of theta from 2000 SNPs, we calculate an ancestral effective population size of 39 217 individuals prior to the split

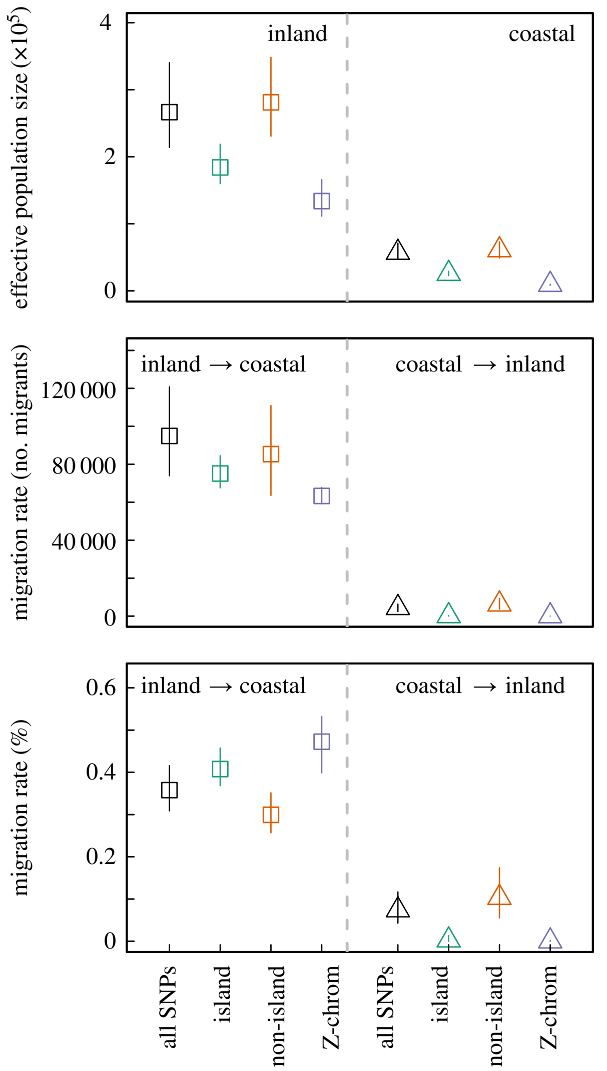

Figure 3. Estimates of effective population size and migration rate for four subsets of SNPs. Points represent median values and error bars reflect 95% CIs.

(s.e. = 507). Current population sizes were significantly greater, based on non-overlapping confidence intervals, for the inland than the coastal subspecies, suggesting a larger demographic expansion in the inland subspecies. Current effective population size for the inland subspecies was estimated to be 570 338 (95% CI = 512 976–641 190) individuals, whereas the current effective size of the coastal subspecies was estimated to be 121 038 (95% CI = 94 467–145 894) individuals.

(b) Heterogeneous gene flow across the genome

We observe a high degree of variation in estimates of population size and migration rate calculated from different genomic regions (figure 3). Here, we present results from 2000 SNP subsamples, but 1000 and 5000 SNP subsamples showed similar patterns (electronic supplementary material, figures S3 and S4). Because we are intentionally violating the neutrality assumptions of demographic models, we do not expect parameter estimates to accurately reflect demographic scenarios, but rather provide relative measures of gene flow between genomic regions. Overall, effective population sizes are higher for subsets of all SNPs and non-island SNPs than for SNPs in islands of divergence and those on the Z chromosome. Not unexpectedly, migration rate, measured as the number of migrants per generation, paralleled effective population size estimates. SNPs on islands and the Z chromosome had lower numbers of migrants than all SNPs or non-island SNPs. Overall, more migration occurs in the direction of inland to coastal subspecies.

Because representing migration as number of migrants is highly influenced by population size, we calculated migration rate as a fraction of the source population (number of migrants/effective population size), the traditional expression of a migration rate parameter. We still find a very strong signature of asymmetric gene flow between subspecies; migration rates from inland to coastal range from 0.3 to 0.47 compared with 0–0.1 in the other direction. For migration from coastal to inland, we find the expected pattern of restricted gene flow in highly differentiated regions; we estimate higher migration rates for all SNPs and non-island autosomal SNPs. In fact, the migration rate was nearly zero for island SNPs (0.002) and Z-chromosome SNPs (1×10^{-5}). Migration in the other direction, however, followed the opposite pattern. Differentiated regions, island and Z-chromosome loci, showed higher levels of gene flow from inland to coastal subspecies than non-island regions. In fact, the lowest migration rate was in non-island autosomal SNPs (0.30) and the highest migration rate was in Z-chromosome SNPs (0.47). It is important to note that we do not believe these migration rates accurately reflect the per-generation number of migrants because the demographic model assumes a single, consistent migration rate since divergence rather than complete allopatry followed by secondary contact, which is the case for this species.

(c) Inverse patterns of nucleotide diversity at highly differentiated loci

Within the inland subspecies, nucloetide diversity (π) was significantly lower at more differentiated ($F_{ST} > 0.1$) loci than at all loci (t-test $p < 0.0001$; figure 4), a signal that is expected if differentiated loci were driven to high frequency via selective sweeps. Surprisingly, highly differentiated loci in the coastal subspecies had higher levels of nucleotide diversity than the all SNPs category (t-test $p < 0.0001$), which could occur if selected loci in the inland subspecies were preferentially moving across hybrid zone boundaries. While mean nucleotide diversity in island regions was lower overall within both ecotypes [17,22], this pattern is driven by low F_{ST} loci. Separating the within-island data into low and high F_{ST} SNPs allowed us to specifically examine diversity for the most differentiated loci. These patterns of nucleotide diversity also exist in non-island SNPs, suggesting this asymmetry is not strictly associated with island regions.

(d) Interactions between selection, divergence time, and gene flow

We used forward simulations to further test the impacts of timing of secondary contact, selection, recombination and gene flow on generating the patterns of differentiation and diversity observed. Electronic supplementary material, table S1 shows mean and maximum F_{ST} as well as expected heterozygosity for both island and non-island regions, and a summary is shown in table 2. Here we discuss the probability that the following simulated scenarios generated patterns observed across the Swainson's thrush genome: (H1) neutral mutation only, (H2) selective sweeps in both subspecies during allopatry, (H3a) selective sweeps in the inland subspecies, where alleles under

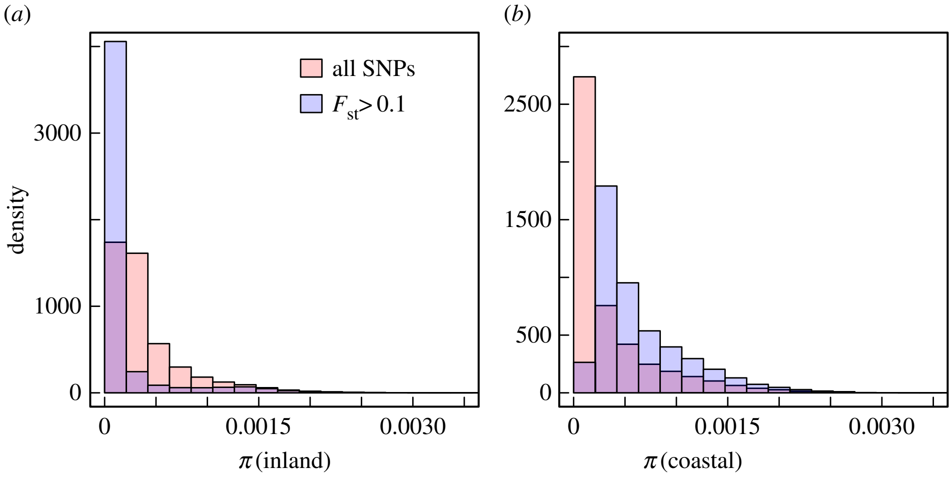

Figure 4. Comparison of nucleotide diversity for all SNPs and differentiated SNPs ($F_{ST} > 0.1$) in the inland (*a*) and coastal (*b*) ecotype of Swainson's thrush. Purple represents the intersection of the two distributions. In the inland ecotype, differentiated regions have lower nucleotide diversity, whereas in the coastal ecotype, differentiated regions have increased nucleotide diversity.

Table 2. Summary of forward simulation results. Observations are patterns generated from real Swainson's thrush genomic data. The table shows whether these patterns were also generated under the three broad scenarios we simulated: (i) neutral mutation only, (ii) selective sweeps occurring in both populations during allopatry and (iii) selective sweeps in inland subspecies only during allopatry.

	highly differentiated (max $F_{ST} > 0.4$)	counterintuitive patterns of nucleotide diversity
H1. neutral mutation only	very recent (100 years) contact only	no
H2. selective sweeps in both subspecies	very recent (100 years) contact only	no
H3a. selective sweeps in inland only: mutations neutral in coastal	yes	yes
H3b. selective sweeps in inland only: mutations beneficial in coastal	no	no

positive selection in the inland subspecies are neutral in the coastal subspecies and (H3b) selective sweeps in the inland subspecies followed by adaptive introgression.

Although other studies have executed much more thorough simulations testing the role of recombination in the formation of differentiated genomic regions [11], our simulations provide some insights into how these islands arose in our system. Under a neutral model (H1), we did not find that low recombination regions were consistently more differentiated. When selective sweeps occurred in both subspecies (H2), we did see higher differentiation in low recombination regions, but only under recent contact (100 years), likely because highly beneficial alleles are quickly fixed upon secondary contact. When selective sweeps occurred in the inland population only, we observed higher differentiation in low recombination regions under lower migration rates (m1 = 0.035 or 0.0035, m2 = 0.007 or 0.0007), but largely under scenarios where selected alleles were neutral in the coastal subspecies (H3a).

A major outlying question is which conditions generate highly differentiated loci? Under neutral mutation alone (H1), regions of high differentiation were achieved under recent contact (100 years max $F_{ST} = 0.86–0.97$ under different migration scenarios), but not if secondary contact commenced 6000 years ago (max $F_{ST} = 0.238–0.331$). Because it is highly unlikely that secondary contact was as recent as 100 years ago, these findings further support the argument that islands of divergence are likely not the product of neutral processes alone. When strong selective sweeps were simulated in both populations (H2), high F_{ST} values were only achieved under recent contact (100 years: max $F_{ST} = 0.687–0.931$; 6000 years $F_{ST} = 0.254–0.381$), because beneficial alleles quickly sweep to fixation after introgression. The other common feature of genomic islands of divergence, reduced diversity, occurred in some simulations with selective sweeps in both subspecies (H2) or the inland subspecies only (H3a and H3b), but was not consistent across either scenario.

A surprising pattern seen in the Swainson's thrush genomic data is the counterintuitive pattern of nucleotide diversity at differentiated loci. We observed that in the inland population, nucleotide diversity was lower at highly differentiated loci, whereas in the coastal population, the opposite pattern existed. We hypothesize that this increased diversity is the result of the addition of inland alleles into the coastal genepool. We see this counterintuitive pattern of nucleotide diversity only when selective sweeps are simulated within the inland subspecies only and those alleles are neutral within the coastal subspecies (table 2).

While our empirical observations are most consistent with a model of selective sweeps in the inland population followed by neutral introgression into the coastal subspecies (H3a), multiple evolutionary processes almost certainly have shaped the genomic patterns. One potential explanation is that selective sweeps occurred in both lineages during allopatry, causing the decreased heterozygosity in island regions, but that the inland subspecies experience more beneficial mutations potentially due to its larger populations size. This, paired with asymmetric gene flow, results in high levels of introgression in high F_{ST} regions with little or no adaptive advantage in the coastal subspecies.

4. Discussion

The ability to use genome scans to uncover genomic regions involved in adaptation and speciation has yielded advancements in our understanding of evolutionary processes [2,4,9,17]. Across many systems, these types of studies have been used to identify genes and genomic regions thought to be involved in speciation. Often, the most differentiated regions are interpreted as regions most affected by divergent selection pressure [3,6] or involved in reproductive isolation [2]. Here we examine evolutionary processes associated with genomic differentiation between two Swainson's thrush subspecies. We find that the differentiated loci were most likely formed by strong selective sweeps in allopatry, facilitated by a large population expansion in a single subspecies, rather than divergent selection in both subspecies with gene flow. Furthermore, contrary to the idea that genomic islands are focal points for reproductive isolation, our results suggest the most differentiated loci move freely across subspecies boundaries. In addition, strong unidirectional introgression is found both within and outside of previously defined islands of divergence. Overall, these results suggest that while there is a higher density of highly differentiated SNPs within islands, similar evolutionary forces are acting on highly differentiated SNPs, irrespective of island boundaries.

(a) Divergence time suggests non-neutral formation of islands

A neutral explanation for islands of divergence is that drift led to divergence in allopatry and gene flow during secondary contact homogenized non-island regions. One way to test this model is to examine divergence times, as this model requires ancient divergence in order to allow drift to result in genome differentiation. We estimate a divergence time of approximately 64 000 years ago followed by a large degree of expansion in both subspecies. Population size estimates for the coastal subspecies are 3.1 times the ancestral population size, but the inland subspecies had even more rapid growth, with a 14.5-fold increase. Our genome-wide estimates align very closely with previous work using an mtDNA marker and distribution modelling [19], which calculated a two- to threefold increase in the coastal subspecies and a six- to 12-fold increase in the inland subspecies since the last glacial maximum. This recent divergence and rapid expansion provides expectations about the origin of polymorphisms and divergence across the Swainson's thrush genome. First, the relatively recent split between the groups suggests that low genomic differentiation in interisland regions can be better explained by polymorphism segregating within the ancestral population than recent swamping owing to gene flow after secondary contact. Second, given the Late Pleistocene divergence estimates, the high levels of differentiation and low levels of nucleotide diversity within island regions cannot be explained by neutral processes alone. Finally, our forward simulations based on neutral mutation processes only did not achieve the level of differentiation observed in the real data. Rather, our results suggest that selective processes (either adaptive or background selection) were involved in creating these differentiated regions, with hitchhiking promoting lower diversity in regions surrounding clusters of differentiated loci. While background selection is a possibility, corresponding increases in population size provide a higher likelihood of beneficial mutations arising de novo [30] and range expansions can rapidly increase allele frequencies via 'surfing'—the phenomenon whereby adaptive alleles in a rapidly expanding population receive a greater advantage of being adaptive and arriving first to a new territory [31]. Demographic processes such as range expansion can also lead to differentiation of neutral alleles via allele surfing, as drift becomes stronger at the expansion front owing to small population sizes [32]. In the case of the inland subspecies, the nearly fivefold larger population size increase estimated here and an equally large range expansion as estimated by Ruegg *et al.* [19] support the idea that selective sweeps and perhaps surfing were prominent features in the evolution of genomic islands of divergence, particularly in the inland subspecies. Finally, the hypothesis that genomic islands are driven by selection in allopatry is consistent with the finding that absolute divergence (d_{xy}) is not lower within differentiated islands [17,22].

(b) Patterns of gene flow and nucleotide diversity support adaptive introgression rather than reproductive isolation

Across the genome, we measured highly asymmetric gene flow, with substantially higher migration rates from inland to coastal populations. Previous studies have found asymmetry in hybridization events, using genotyping of hybrid individuals [33]. This was hypothesized to be the result of migration timing, where migrants of the coastal subspecies arrive on the breeding grounds earlier than the inland subspecies, which migrate longer distances. However, an equally plausible and complementary explanation for the general asymmetry in patterns of gene flow is the 4.7-fold difference in population sizes, as migration predominantly occurs from a larger population to a smaller population, particularly if overall differences in population size are mirrored within the hybrid zone [34].

In addition to asymmetric gene flow between the two subspecies, we find that gene flow is heterogeneous across the genome. Ruegg *et al.* [17] identified 132 clusters of high F_{ST} SNPs, which they called genomic islands of divergence and these results were later corroborated using whole-genome pooled re-sequencing [22]. Explicit demographic modelling allowed us to test expectations under three separate models for the formation and maintenance of differentiated regions (figure 1). Above, we ruled out a neutral model, because the divergence time we estimated was too recent for drift to have differentiated the genomes. We use further estimates of gene flow and nucleotide diversity to test whether patterns of genome-wide differentiation fit the expectations of (i) a divergence with gene flow model in which nucleotide diversity and gene flow within islands is expected to be low or (ii) a selection in allopatry model in which evidence for selective sweeps (high differentiation and low nucleotide diversity) is not necessarily symmetrical. Under a selection with allopatry model, the level of gene flow following secondary contact could reflect selection on alleles on new genetic backgrounds and in new environmental contexts as they move across the hybrid zone.

We propose a mechanism whereby natural selection within the inland subspecies is followed by introgression of potentially adaptive alleles into the coastal subspecies upon secondary contact. Within the inland subspecies alone, we see the classic signals of selective sweeps: high differentiation accompanied by low diversity. In addition, high population sizes within the inland population would provide more opportunity for beneficial mutations to arise [30]. Gene flow from inland to coastal subspecies is actually highest within island and Z-chromosome regions, which is contrary to our expectations under a model of divergent selection with gene flow. The levels of gene flow in this direction within differentiated regions are even higher than genome-wide estimates, which could indicate that these alleles are also beneficial within the coastal group. Forward simulations, however, suggest that selection on these alleles within the coastal subspecies must be small or neutral in order to account for the maintenance of differentiated loci. However, it is possible that introgression of highly adaptive alleles occurred and that those alleles were immediately fixed upon secondary contact. In addition, the high nucleotide diversity we observe in the coastal subspecies at differentiated SNPs could be the result of an influx of inland alleles. It is important to note that this pattern was observed genome-wide and not exclusively within island regions. We suggest that the high gene flow estimates within island regions are a function of an increased fraction of these differentiated loci, but that island regions are not exclusively involved in selection.

This highly asymmetric sharing of genes does raise questions about mechanisms maintaining species boundaries in this system. Our forward simulations suggest that while we have strong support for selection in allopatry, largely in the inland subspecies, other more complex demographic and evolutionary processes contribute to the origin and maintenance of differentiation. One possibility is that the majority of highly differentiated loci are mildly adaptive and moving across the hybrid zone boundary, but a smaller number of loci important to reproductive isolation are failing to move. While we did not have the dataset necessary to test this idea, future studies might make use of whole genome sequencing and cline analysis across the hybrid zone to identify ‘speciation loci’ and assess whether movement is particular to some regions more than others. Alternatively, it is also possible that the range of the inland subspecies is still expanding and may eventually take over the range of the coastal subspecies, leading to extinction of the coastal genotype. While this would run counter to previous expectations and data suggesting consistency in the location of the hybrid zone over the last several decades [19], a more thorough resurvey of the hybrid zone could be used to test the hybrid zone movement hypothesis.

5. Conclusion

With the increasing ease of genomic sequencing in natural populations, we have unprecedented power to understand the processes that lead to speciation. We show that explicit demographic models based on different genomic regions can elucidate evolutionary mechanisms at a level absent from the commonly used F_{ST} scans. In the Swainson’s thrush system, we have the opportunity to witness the early stages in the process of divergence where two groups are highly differentiated across large portions of their genome, but are still exchanging genes at points of contact. In this brief window of time, we find evidence that the most differentiated loci likely arose via selection at the leading edge of a rapid population expansion in one subspecies, followed by high levels of introgression of these potentially adaptive alleles into the other ecotype. Finally, although differentiated loci are clustered into islands of divergence, the patterns of selection and introgression are not restricted to these islands. This extra level of information has increased our understanding of evolutionary forces in this system and provided a more nuanced view of islands of divergence.

Data accessibility. The code for simulations can be found at http://github.com/rachaelbay/SWTH-demography. No new data were generated for this study.

Authors’ contributions. K.C.R. conceived of the study. K.C.R. and R.A.B. designed the study. R.A.B. carried out the analysis. K.C.R. and R.A.B. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. This work was made possible thanks to a generous donation from Margery Nicolson and an NSF postdoctoral fellowship supporting R.A.B.

Acknowledgements. We gratefully acknowledge Ying Zhen, Kelly Barr, Eric Anderson, Sonya Clegg, and Thomas Smith the helpful comments.

References

1. Mayr E. In press. *Systematics and the origin of species, from the viewpoint of a zoologist*. Cambridge, MA: Harvard University Press.
2. Turner TL, Hahn MW, Nuzhdin SV. 2005 Genomic islands of speciation in *Anopheles gambiae*. *PLoS Biol.* **3**, e285–e287. (doi:10.1371/journal.pbio.0030285)
3. Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. 2010 Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet.* **6**, e1000862. (doi:10.1371/journal.pgen.1000862)
4. Nadeau NJ *et al.* 2011 Genomic islands of divergence in hybridizing *Heliconius* butterflies identified by large-scale targeted sequencing. *Phil. Trans. R. Soc B* **367**, 343–353. (doi:10.1098/rstb.2011.0198)
5. Cruickshank TE, Hahn MW. 2014 Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* **23**, 3133–3157. (doi:10.1111/mec.12796)
6. Via S. 2011 Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Phil. Trans. R. Soc. B* **367**, 451–460. (doi:10.1098/rstb.2011.0260)
7. Smith TB, Wayne RK, Girman DJ, Bruford MW. 1997 A role for ecotones in generating rainforest biodiversity. *Science* **276**, 1855–1857. (doi:10.1126/science.276.5320.1855)
8. Orr MR, Smith TB. 1998 Ecology and speciation. *Trends Ecol. Evol.* **13**, 502–506. (doi:10.1016/S0169-5347(98)01511-0)
9. Feder JL, Egan SP, Nosil P. 2012 The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350. (doi:10.1016/j.tig.2012.03.009)
10. Nachman MW, Payseur BA. 2011 Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Phil. Trans. R. Soc. B* **367**, 409–421. (doi:10.1098/rstb.2011.0249)
11. Yeaman S, Aeschbacher S, Bürger R. 2016 The evolution of genomic islands by increased establishment probability of linked alleles. *Mol. Ecol.* **25**, 2542–2558. (doi:10.1111/mec.13611)
12. Burri R *et al.* 2015 Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res.* **25**, 1656–1665. (doi:10.1101/gr.196485.115)

13. Grant PR, Grant R, Markert JA, Keller LF, Petren K. 2004 Covergent evolution of Darwin's finches caused by introgressive hybridization and selection. *Evolution* **58**, 1588–1599. (doi:10.1111/j.0014-3820.2004.tb01738.x)
14. Grant PR, Grant R. 2016 Introgressive hybridization and natural selection in Darwin's finches. *Biol. J. Linn. Soc.* **117**, 812–822. (doi:10.1111/bij.12702)
15. Ellegren H *et al.* 2012 The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**, 756–760. (doi:10.1038/nature11584)
16. Poelstra JW *et al.* 2014 The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science* **344**, 1410–1413. (doi:10.1126/science.1253226)
17. Ruegg K, Anderson EC, Boone J, Pouls J, Smith TB. 2014 A role for migration-linked genes and genomic islands in divergence of a songbird. *Mol. Ecol.* **23**, 4757–4769. (doi:10.1111/mec.12842)
18. Mack DE, Yong W. 2000 Swainson's thrush (*Catharus ustulatus*), no. 540. In *The birds of North America*. Philadelphia, PA: Academy of Natural Sciences.
19. Ruegg KC, Hijmans RJ, Moritz C. 2006 Climate change and the origin of migratory pathways in the Swainson's thrush, *Catharus ustulatus*. *J. Biogeogr.* **33**, 1172–1182. (doi:10.1111/j.1365-2699.2006.01517.x)
20. Ruegg K. 2008 Genetic, morphological, and ecological characterization of a hybrid zone that spans a migratory divide. *Evolution* **62**, 452–466. (doi:10.1111/j.1558-5646.2007.00263.x)
21. Ruegg KC, Smith TB. 2002 Not as the crow flies: a historical explanation for circuitous migration in Swainson's thrush (*Catharus ustulatus*). *Proc. R. Soc. B* **269**, 1375–1381. (doi:10.1098/rspb.2002.2032)
22. Delmore KE, Hübner S, Kane NC, Schuster R, Andrew RL, Câmara F, Guigó R, Irwin DE. 2015 Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Mol. Ecol.* **24**, 1873–1888. (doi:10.1111/mec.13150)
23. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009 Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* **5**, e1000695-11. (doi:10.1371/journal.pgen.1000695)
24. Nadachowska-Brzyska K, Li C, Smeds L, Zhang G, Ellegren H. 2015 Temporal dynamics of avian populations during pleistocene revealed by whole-genome sequences. *Curr. Biol.* **25**, 1375–1380. (doi:10.1016/j.cub.2015.03.047)
25. Li H, Durbin R. 2011 Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496. (doi:10.1038/nature10231)
26. Lunter G, Goodson M. 2011 Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* **21**, 936–939. (doi:10.1101/gr.111120.110)
27. Li H *et al.* 2009 The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079. (doi:10.1093/bioinformatics/btp352)
28. Messer PW. 2013 SLiM: simulating evolution with selection and linkage. *Genetics* **194**, 1037–1039. (doi:10.1534/genetics.113.152181/-/DC1)
29. Backstrom N *et al.* 2010 The recombination landscape of the zebra finch *Taeniopygia guttata* genome. *Genome Res.* **20**, 485–495. (doi:10.1101/gr.101410.109)
30. Hartl DL, Clark AG. 1997 *Principles of population genetics*. Sunderland, MA: Sinauer Associates.
31. Excoffier L, Ray N. 2008 Surfing during population expansions promotes genetic revolutions and structuration. *Trends Ecol. Evol.* **23**, 347–351. (doi:10.1016/j.tree.2008.04.004)
32. Hoban S. 2014 An overview of the utility of population simulation software in molecular ecology. *Mol. Ecol.* **23**, 2383–2401. (doi:10.1111/mec.12741)
33. Ruegg K, Anderson EC, Slabbekoorn H. 2012 Differences in timing of migration and response to sexual signalling drive asymmetric hybridization across a migratory divide. *J. Evol. Biol.* **25**, 1741–1750. (doi:10.1111/j.1420-9101.2012.02554.x)
34. Bolnick DI, Nosil P. 2007 Natural selection in populations subject to a migration load. *Evolution* **61**, 2229–2243. (doi:10.1111/j.1558-5646.2007.00179.x)