The Genomic Signature of Allopatric Speciation in a Songbird Is Shaped by Genome Architecture (Aves: *Certhia americana*)

Joseph D. Manthey [1,*], John Klicka[2,3], and Garth M. Spellman[4]

[1]Department of Biological Sciences, Texas Tech University, Lubbock, Texas, USA
[2]Burke Museum of Natural History, University of Washington, Seattle, Washington, USA
[3]Department of Biology, University of Washington, Seattle, Washington, USA
[4]Department of Zoology, Denver Museum of Nature & Science, Denver, Colorado, USA

*Corresponding author: E-mail: jdmanthey@gmail.com.

Accepted: 24 May 2021

Abstract

The genomic signature of speciation with gene flow is often attributed to the strength of divergent selection and recombination rate in regions harboring targets for selection. In contrast, allopatric speciation provides a different geographic context and evolutionary scenario, whereby introgression is limited by isolation rather than selection against gene flow. Lacking shared divergent selection or selection against hybridization, we would predict the genomic signature of allopatric speciation would largely be shaped by genomic architecture—the nonrandom distribution of functional elements and chromosomal characteristics—through its role in affecting the processes of selection and drift. Here, we built and annotated a chromosome-scale genome assembly for a songbird (Passeriformes: *Certhia americana*). We show that the genomic signature of allopatric speciation between its two primary lineages is largely shaped by genomic architecture. Regionally, gene density and recombination rate variation explain a large proportion of variance in genomic diversity, differentiation, and divergence. We identified a heterogeneous landscape of selection and neutrality, with a large portion of the genome under the effects of indirect selection. We found higher proportions of small chromosomes under the effects of indirect selection, likely because they have relatively higher gene density. At the chromosome scale, differential genomic architecture of macro- and microchromosomes shapes the genomic signatures of speciation: chromosome size has: 1) a positive relationship with genetic differentiation, genetic divergence, rate of lineage sorting in the contact zone, and proportion neutral evolution and 2) a negative relationship with genetic diversity and recombination rate.

Key words: speciation, linked selection, allopatry, introgression, gene flow.

Significance

The geographic context of speciation impacts the population genetic processes that contribute to evolutionary divergence between lineages. Here, we investigated the genomic signature of speciation in a songbird that underwent lineage divergence in allopatry with a lack of gene flow between lineages. We found that genetic variation within and between lineages covaried with variation in gene density and recombination rates across the genome, likely due to the interactions between gene density, recombination, and population genetic processes such as natural selection and genetic drift.

Introduction

Geographic isolation plays a vital role in the generation of reproductive isolation between populations. Consequently, allopatric speciation has generally been at the forefront of speciation research since the role of isolation was emphasized by evolutionary biologists during the modern synthesis (Dobzhansky 1937; Mayr 1942). Over the past couple

decades, there has been an increasing appreciation for speciation with gene flow, whereby allopatry is not strictly necessary during the speciation process (Feder et al. 2012). Regardless, speciation's geographic aspect needs consideration because the amount of geographic contact between diverging populations will directly impact when and how much gene flow occurs during the speciation process (Nosil and Feder 2012).

The genomic signature of speciation with gene flow is shaped by variation in strength of divergent selection, patterns of linked selection, and rates of gene flow and recombination (Feder et al. 2012; Nosil and Feder 2012). These patterns naturally vary with the geographic context of speciation. For instance, diverging populations in early stages of speciation will have little reproductive isolation, and if they have extensive geographic contact—for example, sympatric populations or brief isolation followed by secondary contact—their genomes will show contrasting peaks and valleys of genetic differentiation (Ellegren et al. 2012). Peaks (a.k.a. islands) of differentiation form due to the effects and interactions of divergent selection, linked selection, reduced gene flow between populations, and reduced recombination rate in genomic regions that harbor targets for selection (Feder et al. 2012; Cruickshank and Hahn 2014). In contrast, valleys of low background genetic differentiation are caused by a lack of barriers to gene flow between diverging species' genomes. The majority of recent empirical speciation genomics research has involved species in geographic contexts allowing gene flow during early and intermediate stages of reproductive isolation (Ellegren et al. 2012; Gagnaire et al. 2013; Carneiro et al. 2014; Janoušek et al. 2015; Toews et al. 2016; Riesch et al. 2017; Westram et al. 2018). These studies have generally found peaks of differentiation scattered across the genome with an apparently higher abundance in genomic regions with relatively low recombination (Nachman and Payseur 2012; Payseur and Rieseberg 2016).

Despite the recent plethora of studies investigating the genomics of speciation with gene flow, allopatric speciation is widely regarded as the most common mode of speciation (Mayr 1942; Futuyma and Mayer 1980). In contrast to speciation with gene flow, in geographic contexts where population divergence occurs in allopatry, genomic divergence is free to proceed via genetic drift or within-population selection without the homogenizing effects of gene flow. Under these circumstances, some have hypothesized to expect a slow accumulation of background differentiation between lineages without clear extremes of differentiation across the genome (Feder et al. 2012). As such, we might predict that the majority of genomic differentiation is linked to within-lineage background selection and indirect selection in regions with relatively low recombination rates (Renaut et al. 2013; Cruickshank and Hahn 2014; Burri et al. 2015), as opposed to divergent selection or genomic regions with strong selection against gene flow in contact zones. If this is the case, we hypothesize the heterogeneity of genomic architecture—the nonrandom distribution of functional elements (e.g., genes, repetitive elements) (Koonin 2009) and chromosomal characteristics (e.g., variable recombination rates)—largely shapes patterns of genomic variation in species that undergo allopatric speciation.

Vertebrate genomes are highly heterogeneous in structure and content (i.e., genomic architecture), including: 1) chromosome size, 2) gene density, 3) repetitive element density, and 4) local recombination rate. For example, bird genomes have chromosomes that span two orders of magnitude in size (Ellegren 2010). Their smaller microchromosomes are denser in gene content (Dutoit, Burri, et al. 2017), bird sex chromosomes tend to be enriched in repetitive elements (Kapusta and Suh 2017), and local recombination rates can vary within and among chromosomes by an order of magnitude or more (Kawakami et al. 2014). Given this genomic architecture, we expect macro- and microchromosomes to show significantly different patterns of genomic variation in birds that have undergone allopatric speciation without prominent gene flow or strong divergent selection (Dutoit, Burri, et al. 2017).

Here, we use a wild songbird to investigate how genomic architecture shapes the genomic signature of allopatric speciation. The Brown Creeper (*Certhia americana*) is widely distributed in most forested habitats from Alaska to Nicaragua. Its two main lineages largely evolved in allopatry (Manthey et al. 2011a, 2011b, 2014) and meet in a microallopatric contact zone in the sky islands—montane forest habitat islands—of Arizona (Marshall 1956; Manthey et al. 2016). We assembled and annotated a chromosome-scale genome for *C. americana* and resequenced individuals from allopatric and contact zone populations. We estimated patterns of genetic diversity, genetic differentiation and divergence, recombination rate, introgression, lineage sorting, and natural selection across the genome, and linked population genetic patterns with genomic architecture. The sampling of both allopatric populations and populations in secondary contact allows us to better understand how genomic architecture affects genomic divergence under the most common geographic mode of speciation as well as how this architecture can affect evolutionary processes in secondary contact.

Results

Chromosome-Level Assembly

We used three sequencing methods to assemble a *C. americana* genome: 1) Pacific Biosciences long reads (~29× coverage), 2) 10x Chromium sequencing (~51× coverage), and 3) Hi-C sequencing (>1,000× physical coverage). The resulting genome had a scaffold L50 of 64.36 Mb, with 91.2% assembled into the 30 largest scaffolds that were highly syntenous with *Taeniopygia guttata* chromosomes (fig. 1). The assembly consisted of ~10.1% repetitive elements, and contained ~98% of BUSCO (Simão et al. 2015) conserved single-

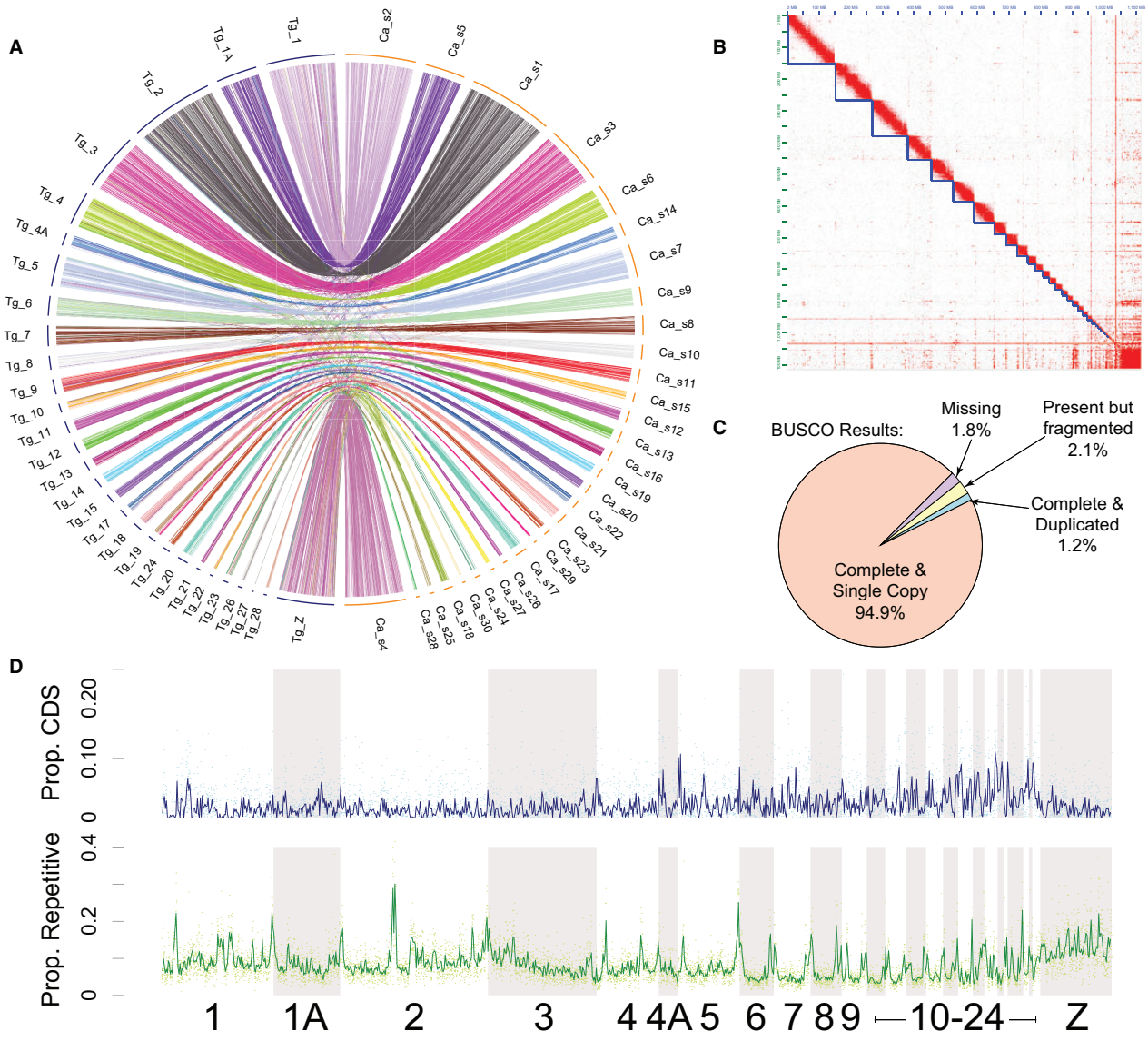

Fig. 1.—Genomic characteristics of the *Certhia americana* de novo assembly. (*A*) Synteny mapping *C. americana* scaffolds (right) to *Taeniopygia guttata* chromosomes (left). (*B*) Hi-C contact map, a heatmap of paired-end Hi-C reads. The blue lines indicate bounding areas for scaffolds. (*C*) Results of BUSCO search for single-copy vertebrate conserved genes in the de novo assembly. (*D*) Proportions of coding sequence (CDS) and repetitive elements in the assembly across 100 kb sliding nonoverlapping windows. The lines indicate mean values across ten sliding windows (i.e., 1 Mb).

copy vertebrate genes (fig. 1). Microchromosomes had denser gene content than macrochromosomes, the Z chromosome showed relatively higher mean repetitive content, and autosomal chromosomes exhibited higher repetitive content on their ends (fig. 1).

Evolutionary Relationships

We resequenced three individuals each for parental populations of both lineages and six populations in the putative contact zone (fig. 2 and supplementary fig. S1 and table S1, Supplementary Material online). To estimate evolutionary relationships and genetic structure, we used a subset of single-nucleotide polymorphisms (SNPs) with no missing data and separated by a minimum of 10 kbp to reduce effects of linkage (supplementary table S2, Supplementary Material online). We estimated a species tree in TreeMix (Pickrell and Pritchard 2012) to infer the proportion of variance in SNP data explained by: 1) independent evolutionary history across lineages and 2) putative gene flow. The base species tree explained ~99.86% of variance in the SNP data, with two putative gene flow events explaining an additional ~0.08% and ~0.06% (fig. 2*D*). With the same SNPs, we used discriminant analysis of principal component (DAPC) (Jombart et al.

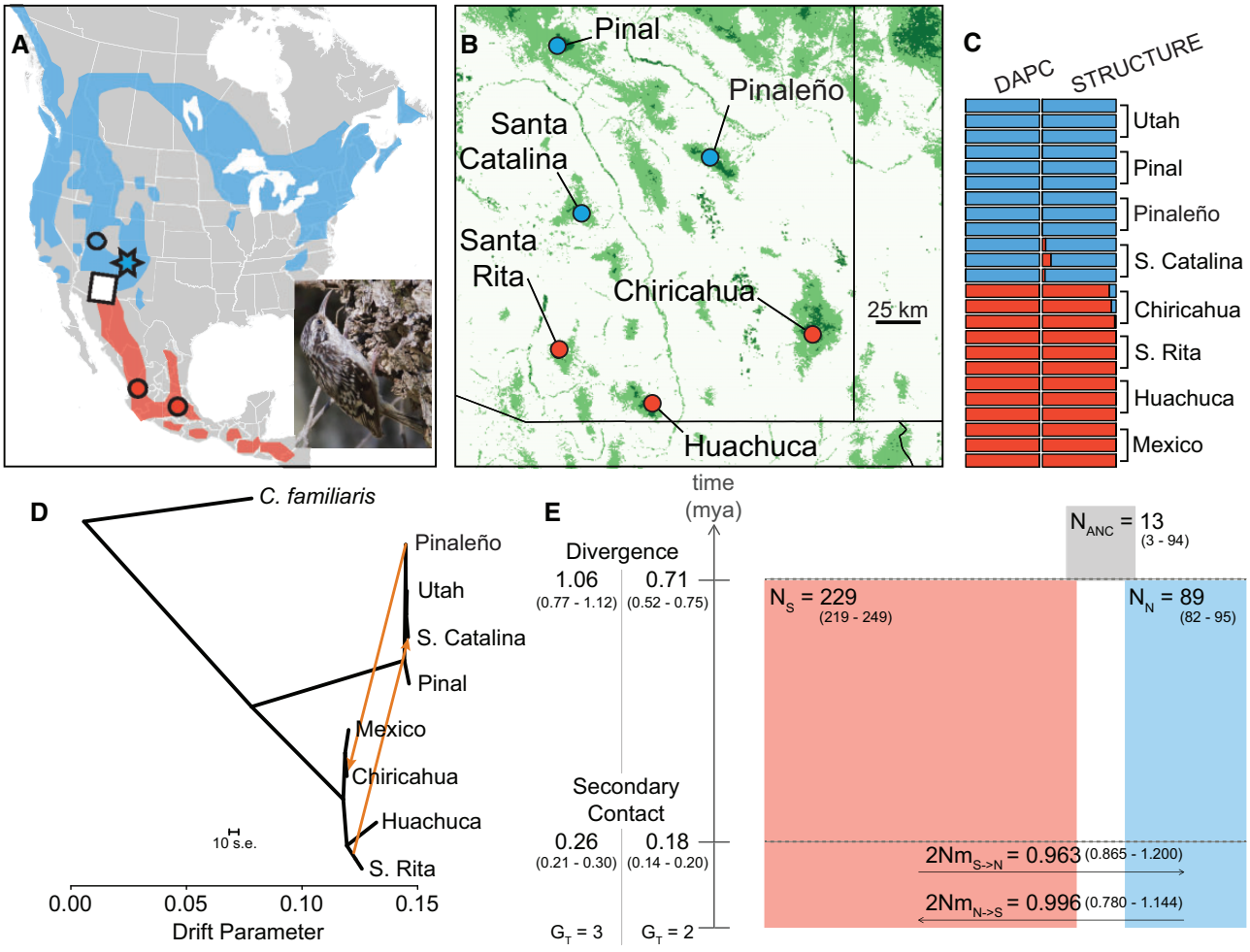

Fig. 2.—Sampling, genetic structure, and demography. (*A*) Map of two main *Certhia americana* lineages. The star indicates locality of individuals used for genome assembly. The inset zooms in on sampling localities in the Arizona sky islands in (*B*) where darker shades of green indicate increased vegetative cover. (*C*) Genetic structure across all individuals inferred in the programs DAPC and STRUCTURE. (*D*) TreeMix species tree of the sampling localities, with two inferred migration edges between lineages. The first is from ancestral Santa Rita Mountains population to the Santa Catalina Mountains, and the second connects migration from the Pinaleño Mountains to the Chiricahua Mountains. Both migration edges explain <0.1% of the variance in the SNP data. (*E*) Demographic history estimated with fastsimcoal2. The best fitting model was identified as a secondary contact model. The times are scaled based on the assumption of 2- or 3-year generation times (G_T). Effective population sizes are in thousands. Confidence intervals estimated from 100 bootstrapped data sets are indicated in parentheses. Photo of Brown Creeper from Chiricahua Mountains taken by J.D.M.

2010) and STRUCTURE (Pritchard et al. 2000) to infer genetic structure; individuals were clearly assigned to one of two lineages (fig. 2*C*). In the STRUCTURE analysis, individuals from two populations (Santa Catalina and Chiricahua Mountains) showed small probabilities of admixture (Q-coefficients <10% for introgressed lineage; fig. 2*C*). Taken together, TreeMix and STRUCTURE analyses showed a small but nonzero signal of introgression, but notably only in two populations and not spread across the entire sampling region.

Population Genomics Summary Statistics

We estimated population genetic summary statistics in 100 kbp sliding windows and mean values per chromosome. At the chromosome level, we found strong negative relationships between chromosome size and within-lineage genetic diversity and recombination rate estimates (fig. 3). These relationships correspond with a strong positive association between chromosome size and both between-lineage genetic differentiation (F_{ST}; $r = 0.885$) and genetic divergence (D_{XY}; $r = 0.839$; fig. 3).

Across sliding windows, estimates of genetic differentiation, genetic divergence, genetic diversity, and recombination rate were highly heterogeneous (fig. 4*A*). Although there were general patterns exhibited based on mean estimates across differently sized chromosomes (fig. 3), these statistics varied greatly within chromosomes (fig. 4*A*). Generally, recombination rate was highly correlated between lineages

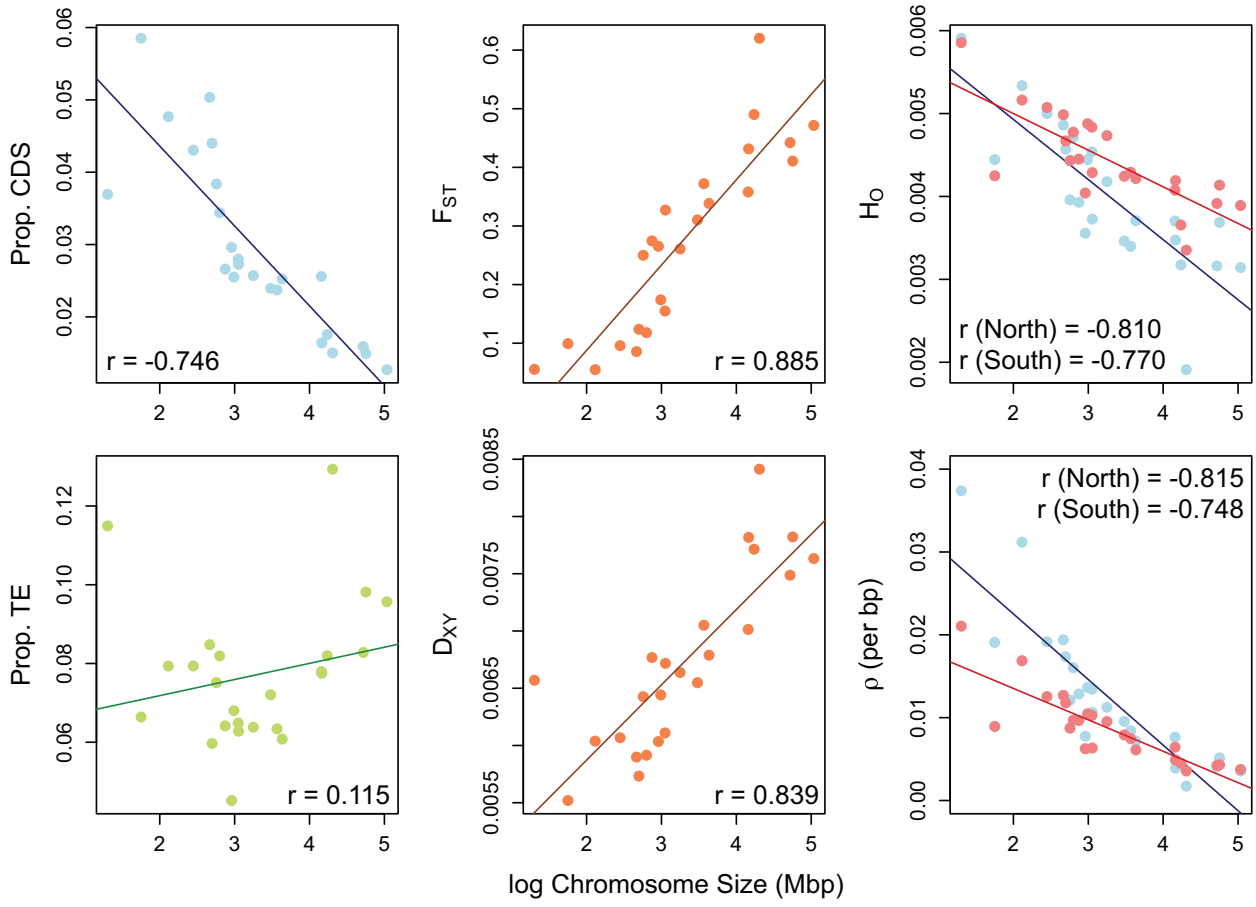

Fig. 3.—Chromosomal mean characteristics. The relationships between chromosome size and mean summary statistics: gene content (CDS), transposable elements (TEs), genetic differentiation (F_{ST}), genetic divergence (also called between lineage diversity; D_{XY}), genetic diversity (H_O), and recombination rate (ρ). In the diversity and recombination plots, red represents southern lineage and blue represents northern lineage.

($r = 0.829$–0.861 across multiple correlation metrics; supplementary table S3, Supplementary Material online) and was positively associated with gene content ($r = 0.241$–0.455). Additionally, regions with high recombination rates—usually chromosome ends and smaller chromosomes—exhibited relatively higher genetic diversity ($r = 0.680$–0.839), lower genetic differentiation ($r = -0.879$ to -0.670), and lower genetic divergence ($r = -0.422$ to -0.162). Notably, the correlation between genomic divergence and 1) genetic differentiation, 2) genetic diversity, and 3) recombination rate all showed a strong relationship with chromosome size (fig. 4*B*), suggestive of different dominant evolutionary forces on different sized chromosomes.

Because we identified correlations between population genomic statistics and chromosomal characteristics at both the chromosomal (fig. 3) and sliding windows (fig. 4*A*) levels, we wanted to parse out if there were any chromosomal effects not accounted for by the gene content, transposable element (TE) content, and local recombination rate in sliding windows. Using partial regression to account for the local window characteristics (e.g., gene content), we found: 1) a positive relationship between chromosomal size and genetic differentiation ($r = 0.496$) and 2) genetic divergence ($r = 0.715$), but also (3) a lack of a relationship between chromosomal size and genetic diversity for either lineage (northern $r = -0.110$, southern $r = 0.170$; supplementary fig. S2, Supplementary Material online). These results indicate that local genomic architecture (e.g., gene and TE content, recombination rate) drives the entire signal between chromosome size and genomic diversity. In contrast, part of the variance in genetic differentiation and divergence is explained by chromosome size even when local genomic architecture is accounted for (supplementary fig. S2, Supplementary Material online). We additionally used partial regression to assess the direct association between recombination rate and population genomic summary statistics at the chromosome scale while accounting for gene content, TE content, and chromosome size. We found strong correlations between recombination rate and genomic diversity (northern $r = 0.673$, southern $r = 0.636$) and between genetic differentiation (F_{ST}) and recombination rate ($r = -0.520$; supplementary fig. S2, Supplementary Material online). In contrast, genetic

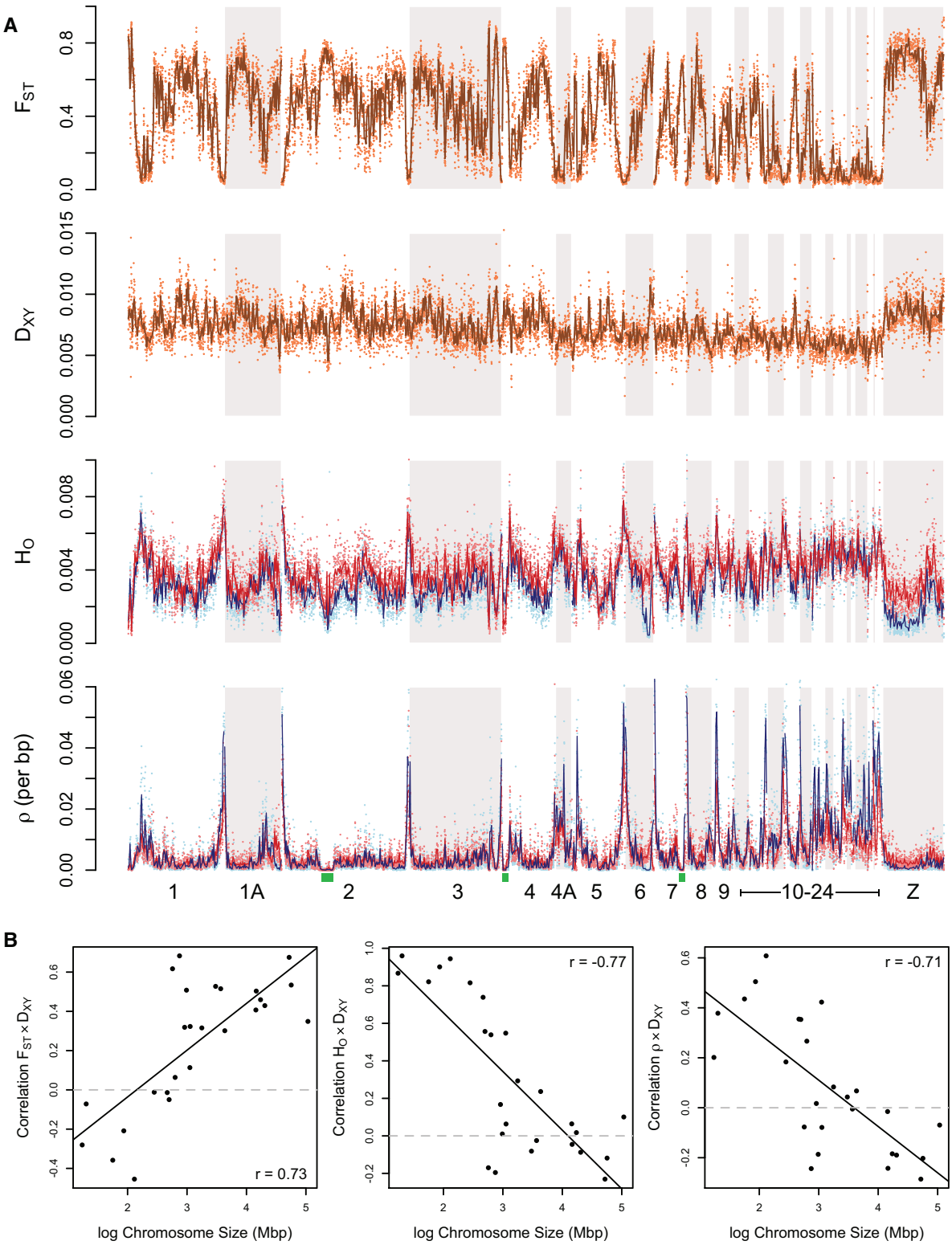

Fig. 4.—(*A*) Population genomic estimates in 100 kbp sliding windows. Population genomic statistics: genetic differentiation (F_{ST}) and divergence (D_{XY}), genetic diversity (H_O), and recombination rate (ρ). The lines indicate mean values across ten sliding windows (i.e., 1 Mb). Green blocks under recombination plot ($N = 3$) indicate regions with putative inversions. (*B*) Chromosome-scale variation in the correlation (Pearson) between genetic divergence (D_{XY}) and other summary statistics across sliding windows. For diversity and recombination measures, red = southern lineage and blue = northern lineage.

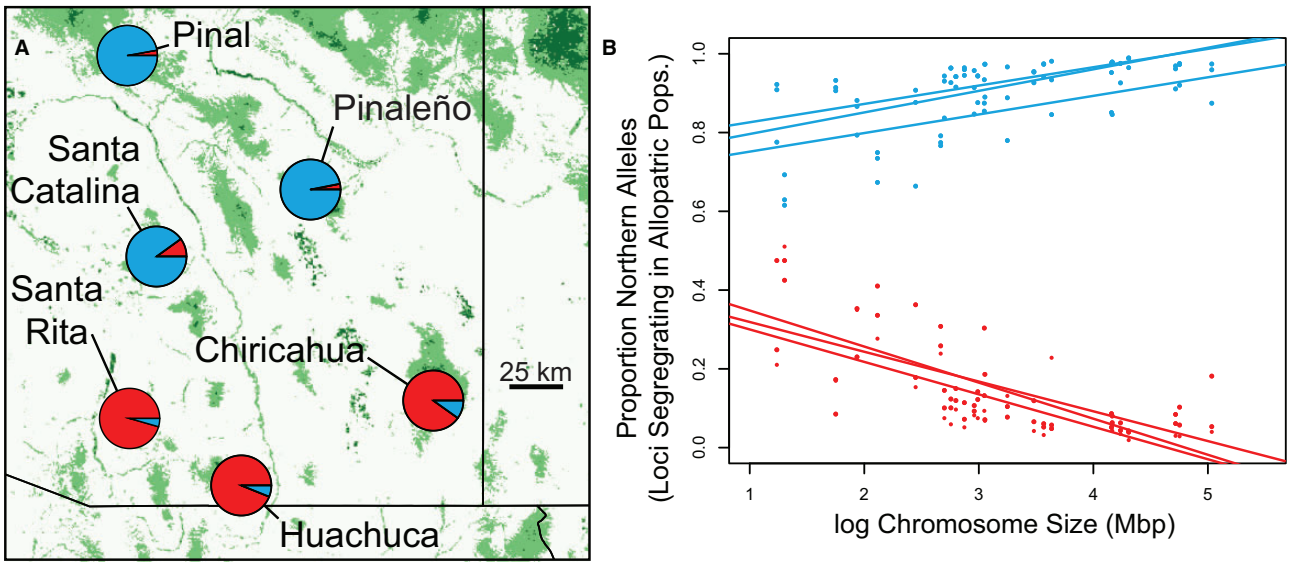

Fig. 5.—Segregating sites in sky island populations of *Certhia americana*. (*A*) Proportion of polymorphisms segregating in allopatric populations found in the sky island populations. (*B*) Segregating site proportions for each chromosome and each population. Regression lines show relationships for each population.

divergence (D_{XY}) and recombination rate were weakly correlated ($r = 0.290$; supplementary fig. S2, Supplementary Material online).

Segregating Sites and Introgression

Because of the potential for some gene flow between lineages (fig. 2), we investigated patterns of: 1) differential fixation across the genome, 2) proportions of sites segregating between allopatric populations found in sky island populations, and 3) introgression using ABBA–BABA tests (both D and f_d statistics; Green et al. 2010; Martin et al. 2015). The proportion of variants fixed in 100 kbp windows (both SNPs and indels) was heterogeneous with some windows having no fixed differences and others showing >50% of polymorphisms fixed between lineages (supplementary fig. S3, Supplementary Material online). Generally, larger chromosomes had higher proportions of fixed differences than smaller chromosomes (supplementary fig. S3, Supplementary Material online). We looked at sites representing fixed differences between the allopatric populations as candidate segregating sites between the lineages, and measured proportions of these sites that were sorting to the northern or southern lineages' alleles in the sky island populations. Here, the sky islands largely exhibited the expected alleles based on their lineage, with relatively higher evidence of allele sharing between lineages in the Santa Catalina and Chiricahua Mountain Ranges (fig. 5*A*). Of these segregating sites, sky island populations showed higher affinity to lineage-specific alleles on larger chromosomes, with smaller chromosomes exhibiting relatively higher segregating-site allele sharing between lineages (fig. 5*B*). These patterns could indicate faster lineage sorting on larger chromosomes, reduced gene flow on larger chromosomes, or both.

For each of the sky island populations, we performed ABBA–BABA tests in 100 kbp sliding windows to identify: 1) whether there was gene flow between lineages in the contact zone, and if so, 2) whether gene flow occurred across the whole genome. Here, we found little evidence of gene flow into most sky island populations, with the exception of the Santa Catalina and the Chiricahua Mountain Ranges (fig. 6). Approximately 6.6% and 3.8% of ~9,800 sliding windows showed significant evidence for gene flow in the Chiricahua and Santa Catalina populations, respectively. In contrast, the other four sky island populations showed evidence for introgression in 0.4–1.5% of windows (fig. 6).

Demography and Selection

We estimated each lineage's demographic history and divergence timing using fastSimCoal2 (Excoffier et al. 2013). Of four alternative models tested in fastSimcoal2, a model of isolation during divergence followed by secondary contact best fit the data (fig. 2*E* and supplementary figs. S4 and S5, Supplementary Material online). Depending on our assumptions about generation times (G_T) in *C. americana*, the two lineages diverged approximately 0.71 Ma ($G_T = 2$ years) or 1.06 Ma ($G_T = 3$ years). They likely came into contact in the last quarter million years, with small amounts of gene flow ($2Nm < 1$) since secondary contact (fig. 2*E*). Consistent with its higher genetic diversity, the southern lineage is estimated to have higher estimates of effective population size relative to the northern lineage (fig. 2*E*).

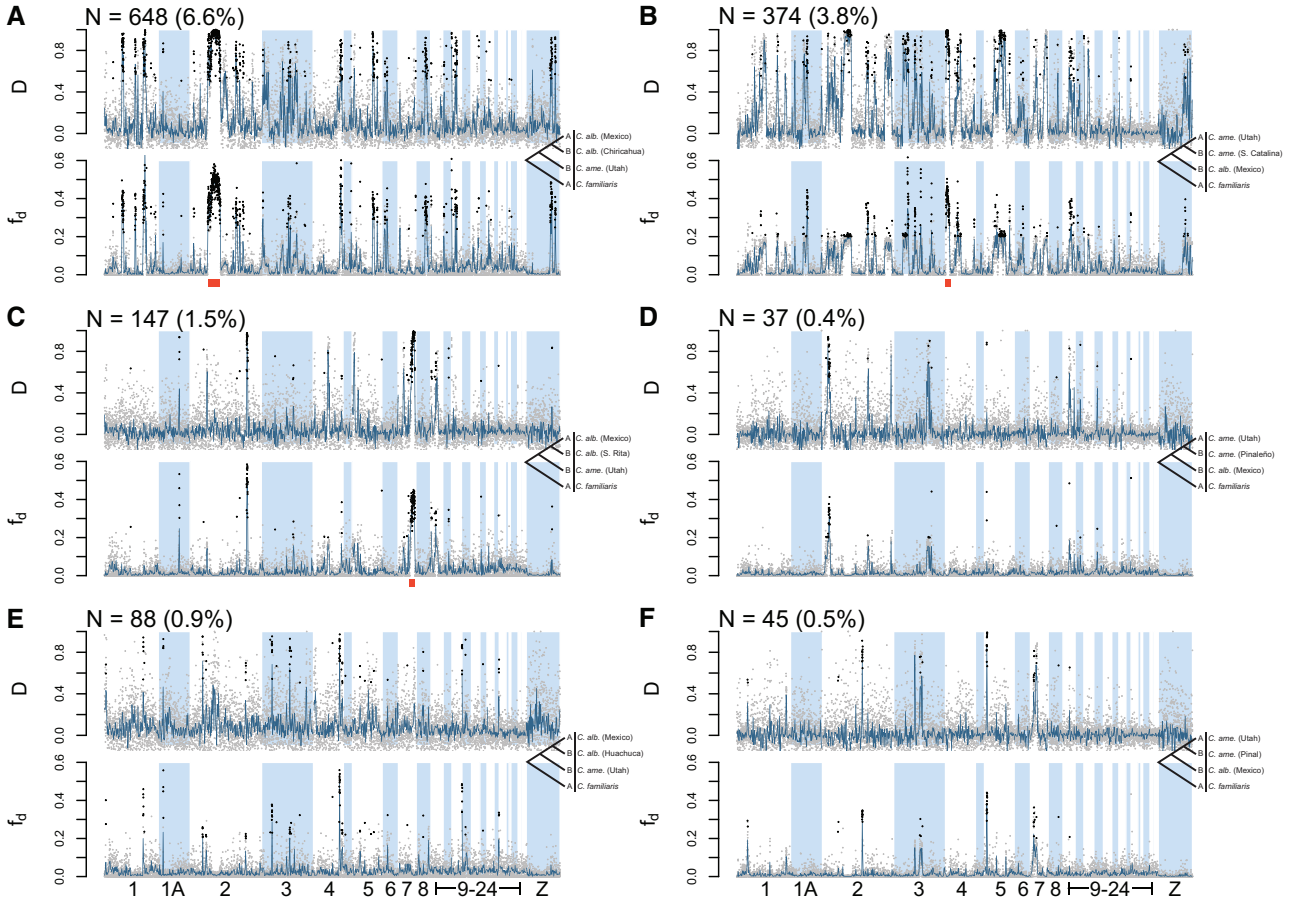

Fig. 6.—Tests for introgression using *D* and f_d statistics. Populations tested: (*A*) Chiricahua, (*B*) Santa Catalina, (*C*) Santa Rita, (*D*) Pinaleño, (*E*) Huachuca, (*F*) Pinal. Gray points show estimates of the statistics per 100 kbp window. Blue lines show 10-window (i.e., 1 Mbp) mean values. Points in windows with $D > 0.5$ and $f_d > 0.3$ are bold, and the count of these bolded points is indicated in the top left of each panel. Red blocks under plots ($N = 3$) indicate regions with putative inversions.

We also estimated demographic histories in smc++ (Terhorst et al. 2017), although this method is likely biased in this case as it does not account for gene flow. Here, we largely wanted to use smc++ to obtain potential fluctuations in effective population sizes through time that could be used as a starting point to account for demographic histories in tests for selection using diploS/HIC (Kern and Schrider 2018). The model estimated in smc++ showed oscillating effective population sizes for both lineages, a divergence time generally consistent with fastSimcoal2 results (~0.87 Ma with $G_T = 2$ years), and harmonic mean effective population sizes larger in the southern lineage relative to the northern lineage (supplementary fig. S6, Supplementary Material online).

We used a machine-learning algorithm implemented in diploS/HIC (Kern and Schrider 2018) to predict patterns of neutrality, selection, and linked selection in 20- and 50 kbp windows. We initially trained the diploS/HIC model using demographically informed coalescent simulations produced with discoal (Kern and Schrider 2016). In both lineages, we found a majority of the genome not evolving neutrally (fig. 7 and supplementary fig. S7, Supplementary Material online). Large portions of the genome were predicted to have undergone soft selective sweeps or be in linkage with soft sweeps (~49–71% of genome), with little evidence of hard sweeps (~1.1–1.4% of genome). Only 14.6% (50 kbp windows) or 13.3% (20 kbp windows) of windows were classified as evolving neutrally in both lineages. A larger proportion of the genome (23–25%) was classified as linked to selection in both lineages. Because of the heterogeneous sizes of chromosomes and their different patterns of coding sequence content and recombination rate, we measured the proportion of windows classified neutral in each chromosome. We found a marginally significant positive relationship between chromosome size and proportion of windows classified neutral (fig. 7; northern $r = 0.419$, southern $r = 0.363$).

Putative Inversion Regions

We identified three regions of the genome greater than 1 Mbp that showed a strong signature of introgression as

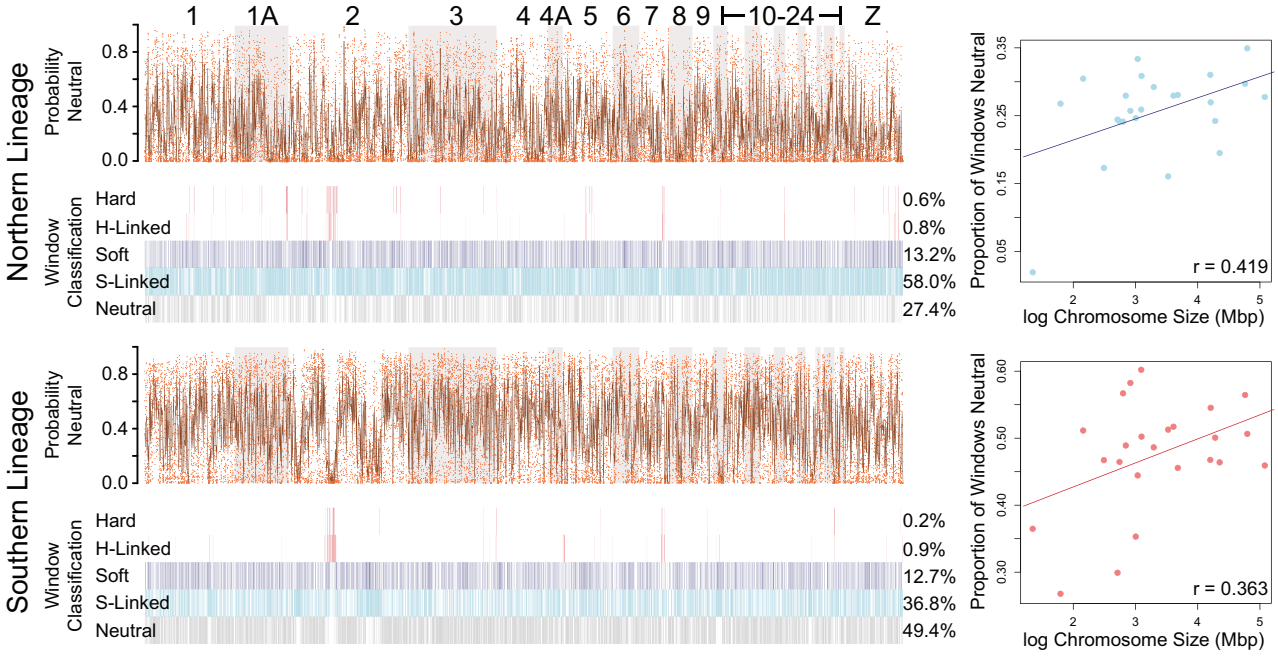

Fig. 7.—Patterns of neutrality and selection. Results of the machine-learning classification for each window inferred with diploS/HIC. We used 50 kbp windows for inference. For each lineage, the top panel shows the probability that each window is neutral (points) with 10-window (i.e., 500 kbp) mean values indicated with lines. The bottom panel for each lineage indicates the classification with the highest probability. On the right, chromosome size is plotted against the proportion of windows on that chromosome classified neutral.

well as reduced recombination rates. We hypothesized that these may be ancestrally segregating inversions and aimed to use multiple lines of evidence to assess this hypothesis. First, we extracted SNPs from these regions with no missing data to estimate genetic structure using the program STRUCTURE. We found that all individuals exhibited either complete ancestry to their respective lineage or approximately half their ancestry to each lineage (i.e., heterozygous for the putative inversion; fig. 8*A*). Individuals with mixed ancestry in these putative inversion regions exhibit nearly 100% heterozygosity of polymorphic sites in these regions (fig. 8*E*). Relative to the chromosomes they are located on, these putative inversion regions exhibit highly reduced recombination rate (fig. 8*B*), decreased heterozygosity (fig. 8*C*), and greatly increased linkage disequilibrium (LD; fig. 8*D* and supplementary fig. S8, Supplementary Material online). Overall, the observed patterns in these three regions are indicative—but not conclusive—of inversions whereby two to three individuals in the contact zone are heterozygous for each of the inversions, but only one individual is heterozygous with all three (fig. 8*A*).

Discussion

Genomic Architecture Shapes the Genomic Signature of Allopatric Speciation in the Brown Creeper

We built and annotated a de novo chromosome-scale genome assembly for *C. americana* to investigate how genomic architecture shapes the genomic landscape of allopatric speciation. The genome's characteristics were highly heterogeneous both within and across chromosomes. We found that smaller chromosomes tended to have higher effective recombination rates (figs. 3 and 4*A*). This trend of relatively higher recombination rates on smaller chromosomes is consistent with patterns identified in other organisms—such as yeasts, birds, butterflies, mammals, and fishes—and is thought to occur due to meiotic crossover requirements (Kaback et al. 1992; Jensen-Seaman et al. 2004; Lynch and Walsh 2007; Roesti et al. 2013; Kawakami et al. 2014; Davey et al. 2017; Martin et al. 2019). Additionally, most chromosomes showed relatively higher recombination rates on chromosome ends (fig. 4*A*), consistent with trends in both plants and animals (Haenel et al. 2018). Recombination rate variation across the genome was generally consistent between the two *Certhia* lineages we examined ($r = 0.829$–0.861 across multiple correlation measures; supplementary table S3, Supplementary Material online).

We found that smaller chromosomes tended to have denser gene content, and in sliding windows gene content was positively correlated with recombination rate (figs. 3 and 4*A*; supplementary table S3, Supplementary Material online), consistent with patterns seen in other songbirds (Kawakami et al. 2014; Dutoit, Burri, et al. 2017). In eukaryotes, recombination rates are generally positively correlated with coding sequence density, but also with notable exceptions demonstrating a negative correlation between local recombination

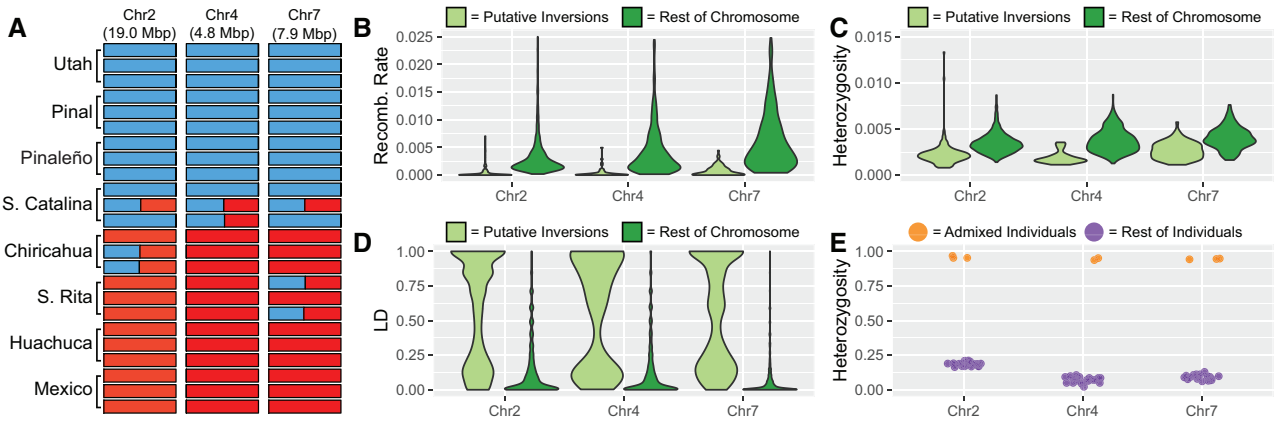

Fig. 8.—Genomic characteristics of putative inversion regions: (*A*) STRUCTURE results, (*B*) recombination rates, (*C*) heterozygosity, (*D*) R^2 estimates of linkage disequilibrium (LD), and (*E*) heterozygosity in individuals admixed/heterozygous for the putative inversions relative to all other individuals (limited to polymorphic sites with MAF $\geq$0.3).

rate and gene content (Stapley et al. 2017; Martin et al. 2019; Stankowski et al. 2019).

The highly heterogeneous genomic landscape of gene content and recombination rates is strongly associated with patterns of genomic diversity, differentiation, and divergence at the chromosome level (fig. 3). Because of the differential genomic characteristics of macro- and microchromosomes, chromosome size explains between 58% and 77% of variation in genetic diversity, differentiation, and divergence at the chromosome scale (fig. 3). When accounting for variation in gene content, TE content, and recombination rate variation across chromosomes using partial regression, chromosome effects no longer explain variation in genomic diversity (supplementary fig. S2, Supplementary Material online), but still explain part of the variation among chromosomes in genetic differentiation (21%) and genetic divergence (49%) (supplementary fig. S2, Supplementary Material online). In contrast to patterns at the chromosome scale identified here: 1) chromosome size and genomic diversity are either decoupled or show a positive relationship in other songbird species (Dutoit, Burri, et al. 2017; Dutoit, Vijay, et al. 2017), and 2) chromosome size and genomic divergence exhibit a negative relationship in some mammals (Tigano et al. 2021), although one study in hummingbirds found a relationship between genomic differentiation and chromosome size (Henderson and Brelsford 2020).

At the regional scale, measured in 100 kbp sliding windows, we found strong correlations of relative recombination rate and genomic architecture—such as gene content—with population genomic statistics including genetic diversity, differentiation, and divergence.

Diversity

Window-based analyses showed a strong positive correlation in genomic diversity between the northern and southern lineages ($r = 0.781$–0.806; supplementary table S3, Supplementary Material online). We also identified strong positive correlations of genomic diversity with both gene content and effective recombination rates in sliding window analyses (fig. 4*A* and supplementary table S3, Supplementary Material online). This pattern is similar to that found in other animals at regional genomic scales (i.e., not chromosomal scale and not very small genomic resolutions) (Kraft et al. 1998; Cutter and Payseur 2003, 2013; Takahashi et al. 2004; Kulathinal et al. 2008; Branca et al. 2011; Roesti et al. 2013; Vijay et al. 2016; Dutoit, Burri, et al. 2017; Dutoit, Vijay, et al. 2017). Additionally, a study in humans showed that the diversity–recombination relationship may be scale dependent, with variable patterns on the scale from one to hundreds of kbp (Spencer et al. 2006).

Differentiation and Divergence

We found strong negative correlations of both genetic differentiation and divergence with gene content and recombination rate in sliding window analyses (fig. 4*A* and supplementary table S3, Supplementary Material online). In other species, window-based analyses have identified a negative relationship between recombination rates and genetic differentiation (Kulathinal et al. 2008; Roesti et al. 2013; Vijay et al. 2016; Stankowski et al. 2019). These results may be attributed to the faster accumulation of genetic differences in low recombination regions due to both the effects of genetic drift and linked selection. In contrast to genetic differentiation, we may not expect a consistent relationship between genetic divergence and recombination rate depending on the speciation scenario. For example, in a speciation model without gene flow, we would expect differentiation (F_{ST}) and divergence (D_{XY}) to be positively correlated and scale with time in isolation. In contrast, during speciation with gene flow, differentiation and divergence may be uncorrelated or

negatively correlated due to the complex interactions between gene flow and selection against gene flow between lineages in different parts of the genome (e.g., Stankowski et al. 2019). Here, our observed positive relationship between genomic divergence and recombination rates is consistent with the isolation model of speciation and also with patterns observed in stonechats (Van Doren et al. 2017).

Multiple Lines of Evidence for an Isolation with Secondary Contact Speciation Model

In songbirds, speciation genomics has focused on speciation with gene flow across many genera: *Ficedula* (Burri et al. 2015), *Corvus* (Poelstra et al. 2014), *Phylloscopus* (Irwin et al. 2016; Lundberg et al. 2017), *Catharus* (Delmore et al. 2015), *Saxicola* (Van Doren et al. 2017), *Vermivora* (Toews et al. 2016), and *Sylvia* (Delmore et al. 2018). In all cases, there is no clear relationship between chromosome size and genetic differentiation, and only ~3% of the variation in genetic differentiation (F_{ST}) is shared across species, whereas ~26% of variation in genomic divergence (D_{XY}) is shared across taxa (Delmore et al. 2018). These patterns suggest the majority of genomic differentiation and divergence is shaped by lineage-specific evolutionary forces during speciation with gene flow, but also that some patterns are driven by shared selective pressures across taxa.

In contrast to speciation with gene flow, here we showed that without the homogenizing effects of gene flow between sister lineages during evolutionary divergence, a large proportion of the variation in genomic differentiation and divergence is shaped by recombination rate variation and genomic architecture. Several lines of evidence support a model of isolation with recent secondary contact between *C. americana* lineages versus a speciation with gene flow model. First, site-frequency spectrum (SFS)-based simulations strongly support the isolation with secondary contact model over a speciation with gene flow model (supplementary figs. S4 and S5, Supplementary Material online). Second, the sky island contact zone populations do not show shared patterns of introgression (fig. 6), as would be expected if gene flow were ongoing during the speciation process, rather than sporadic and low frequency during secondary contact. Lastly, during lineage divergence without gene flow, genomic regions with relatively low recombination rates should accumulate fixed differences faster than genomic regions with high recombination rates. In agreement with this expectation, we find increased proportions of fixed differences—both SNPs and indels—on larger chromosomes relative to smaller chromosomes (fig. 5 and supplementary fig. S3, Supplementary Material online). The accumulation of fixed differences will remove shared standing ancestral variation between lineages and reduce within lineage diversity; as such, regions accumulating more fixed differences (e.g., larger chromosomes) will increase between-lineage values of both genetic differentiation and genetic divergence in the absence of gene flow, leading to a positive correlation between genomic differentiation and divergence (figs. 3 and 4*A*; supplementary table S3, Supplementary Material online). Notably, the positive correlation between D_{XY} and F_{ST} only exists for the largest chromosomes (fig. 4*B*), suggesting smaller chromosomes are relatively more susceptible to a combination of background selection and gene flow. In contrast, under a speciation model with gene flow, we may expect differentiation and divergence to not correlate with one another or to have a negative association, especially when selection—direct or indirect—is involved (Ellegren et al. 2012; Ravinet et al. 2017; Stankowski et al. 2019).

Heterogeneous Genomic Landscape of Neutrality and Selection

In the absence of gene flow, we would expect differentiation and divergence between lineages to occur largely due to lineage-specific selection, background selection, and genetic drift. This is in contrast to speciation with gene flow, where we expect to find the signal of divergent selection or selection against hybridization in contact zones. Here, in both lineages we found a majority of the genome under the influence of direct or background selection (fig. 7). Both lineages showed about 13–14% of the genome under direct selection (signatures of either soft or hard selective sweeps; fig. 7). In *Drosophila*, a study using the diploS/HIC method classified 25–30% of the genome with recent signatures of either hard or soft selective sweeps (Adrion et al. 2020). Only ~3% of the *C. americana* genome showed shared signals of direct selection in both lineages (fig. 7). Even without shared signatures of selection in both lineages, most of the predicted direct selection was on standing variation (i.e., soft sweeps). As such, if this selection was polygenic, it has been argued that selection on standing variation could impact much of the genome and cause correlations of genetic differentiation and genomic architecture (Rockman 2012; Stankowski et al. 2019). We also found between 37% and 58% of the *C. americana* genome under the influence of background selection (fig. 7). Similarly, a recent study using diploS/HIC in chimpanzees and bonobos (genus: *Pan*) found 60–85% of the genome linked to soft selection (Brand et al. 2021). One consideration with our selection results is that our per lineage sample sizes are somewhat low, but also in the same range of haploid genomes as a previous study using diploS/HIC (Brand et al. 2021). The relatively small sample sizes could reduce the effectiveness in confidently classifying regions, particularly differentiating between neutral regions and regions linked with weak soft selection.

Because we found variation in population genomic summary statistics related to chromosome size, we measured the proportion of windows classified as neutrally evolving in each chromosome. We found a marginally significant positive

relationship between chromosome size and proportion of windows classified neutral (fig. 7; northern $r = 0.419$, southern $r = 0.363$). As such, at the chromosomal scale, the proportion of the chromosome under the impacts of soft sweeps or linked to soft sweeps has a negative relationship with genetic differentiation and a positive relationship with both gene density and recombination rates (supplementary fig. S9, Supplementary Material online). Clearly, all of the chromosomes have been impacted by a combination of lineage-specific soft selection (direct selection), indirect selection due to linkage with regions under selection, and genetic drift of neutral variation. However, our data indicate that smaller chromosomes tend to exhibit more signals of selection than larger chromosomes (fig. 7). In general, we may expect greater impacts of linked selection in genomic regions with high genic density as well as longer stretches of linked selection in regions with low recombination rates (Nachman and Payseur 2012; Haenel et al. 2018). The positive correlation between gene density and recombination rate in the *Certhia* genome (supplementary table S3, Supplementary Material online) complicates any simple pattern of linked selection across the genome (fig. 7).

Interestingly, larger chromosomes evolving more neutrally have a positive relationship between genetic differentiation (F_{ST}) and genetic divergence (D_{XY}; fig. 4*B*). The positive correlation between F_{ST} and D_{XY} is consistent with both isolation (e.g., resistance to gene flow) and lineage sorting and fixation of ancestral polymorphisms (Guerrero and Hahn 2017; Han et al. 2017). These results suggest that the combined effects of selection and drift on larger chromosomes have greatly reduced their diversity—especially ancestral diversity—and the continued lack of gene flow on these chromosomes leads to high values of both F_{ST} and D_{XY}.

In contrast, the smallest chromosomes are relatively more under the effects of continued indirect selection (fig. 7) and exhibit a negative correlation between genetic divergence and genetic differentiation (fig. 4*B*). The negative correlation between D_{XY} and F_{ST} is consistent with long-term linked selection acting prior to and during speciation (Nachman and Payseur 2012; Vijay et al. 2017). Additionally, the varying relationships between D_{XY} and other summary statistics across differently sized chromosomes (fig. 4*B*) are consistent with simulations showing that D_{XY} is negatively correlated with the intensity of background selection (Phung et al. 2016; Matthey-Doret and Whitlock 2019). Overall, the genomic architecture and variable recombination rates between large and small chromosomes have driven a highly heterogeneous landscape of lineage divergence via both selection and neutrality in the Brown Creeper. This is generally consistent with the idea that at moderate to late stages in speciation, the genomic landscape of differentiation will largely be shaped by regional variation in recombination rates interacting through time with linked selection (Roesti et al. 2013; Burri 2017).

Three Putative Inversions Exhibit Reduced Recombination and Diversity

We identified three genomic regions greater than 1 Mbp that exhibited the characteristics of an inversion structural variant (fig. 8): 1) high LD, 2) very low inferred recombination rates, 3) reduced heterozygosity, and 4) near 100% heterozygosity in individuals we inferred to be heterozygous for the putative inversions. Notably, these regions showed strong signatures of gene flow in introgression tests (fig. 6). Given the moderate coverage, short-read nature of the data we collected for this study, it is difficult to infer direct breakpoints of putative inversions or definitively claim these are true structural variants. Regardless, based on the genomic signatures in these regions, we hypothesize these regions truly are inversions (fig. 8). In passerine birds, large inversions are common between species, present but less common within species, and are associated with reproductive isolation (Hooper and Price 2017; Hooper et al. 2019).

Two alternative scenarios could lead to the patterns identified in the putative inversions: 1) introgression of the inversions resulting from gene flow in secondary contact, or 2) segregating ancestral polymorphism that gives the signal of introgression due to heterozygous individuals exhibiting the signature of an F1 hybrid in these regions that strongly differs from the rest of those individuals' genomic background. We believe the latter scenario of a segregating ancestral polymorphism to be more plausible; the pattern is consistent with simulations of an inversion with divergent selection across an environmental gradient (Faria et al. 2019). In this context, we would expect the inversion to be quickly fixed away from the environmental gradient, with only populations within or adjacent to the environmental transition to exhibit inversion polymorphism (Faria et al. 2019). The *C. americana* contact zone occurs in southern Arizona, USA, in the environmental transition between temperate and subtropical forests (Wade et al. 2003); *C. americana* inhabits pine, fir, and spruce forests in the Rocky Mountains (northern lineage) that transition to the mixed pine-oak woodlands of the Sierra Madre mountain ranges in Mexico (southern lineage). This environmental transition between forest types coincides with biogeographic transition zones in both mammal and plant communities between the Rocky Mountains and the Sierra Madres (Lomolino and Davis 1997; Wade et al. 2003; Lomolino et al. 2006).

Notably, 16 genomic windows showed evidence of hard selective sweeps in both lineages (fig. 7), with all but one of these windows associated with the putative inversions. These tests help support the hypothesis that these regions are associated with divergent selection, but are also confounded because there is a potential bias in neutral inversions or low recombination regions to identify false positives indicative of natural selection (Lotterhos 2019). To definitively confirm that these genomic regions are putative inversions under the effects of divergent selection, we would need to: 1) confirm

these regions are inversions and 2) sample a transect north to south across the contact zone region to identify whether the polymorphisms have strong clinal patterns associated with the environmental gradient.

Conclusions

We investigated the genomic signature of speciation when lineages diverged largely in allopatry. We showed that without the homogenizing effects of gene flow between sister lineages during divergence, a large proportion of the variation in genetic diversity, differentiation, and divergence is shaped by genomic variation in gene density and recombination rates. Variation of gene density and recombination rates on differently sized chromosomes leads to chromosome-size associations with genetic variation. Comparative genomic studies in additional taxa that speciated allopatrically would provide context as to shared speciation patterns when speciation largely lacks gene flow.

Materials and Methods

Genome Assembly

We used two *C. americana* individuals from New Mexico for genome assembly (supplementary table S1, Supplementary Material online), with one individual used for 10x Chromium sequencing and the other for PacBio and Hi-C sequencing because of limitations on tissue sampling from this small songbird.

We obtained 10x Chromium library sequencing (~51× raw coverage following quality control) using services from HudsonAlpha Institute for Biotechnology (Huntsville, AL). They performed high-molecular weight DNA isolation, quality control, 10x Chromium library preparation, and shotgun sequencing on one lane of an Illumina HiSeqX.

We used services from RTL Genomics (Lubbock, TX) to obtain long-read sequencing data. They performed high-molecular weight DNA isolation using Qiagen (Hilden, Germany) high-molecular weight DNA extraction kits, PacBio SMRTbell library preparation, size selection using a Blue Pippin (Sage Science), and sequencing on four Pacific Biosciences Sequel SMRTcells 1M v2 with Sequencing 2.1 reagents, resulting in ~29× coverage.

To obtain long-range contact information of genome structure, we performed Hi-C sequencing (Dudchenko et al. 2017). We used the Phase Genomics Proximo (Seattle, WA) Hi-C Animal Kit and the manufacturer protocol to create a Hi-C library suitable for sequencing on an Illumina sequencer. We then sequenced the library on a partial Illumina NovaSeq6000 S1 flow cell lane at the Texas Tech University Center for Biotechnology and Genomics. We obtained >1,000× physical distance coverage of the Hi-C data after deduplication.

We assembled the *C. americana* genome in six stages. First, we used the supernova assembler v2.0.1 (10× Genomics, Pleasanton, CA) to create a de novo assembly of the 10x Chromium sequencing reads. Second, we used Canu v1.7.1 (Koren et al. 2017) to de novo assemble the Pacific Biosciences long sequencing reads. Third, we used LINKS v1.8.6 (Warren et al. 2015) to merge the 10x Chromium and long-read assemblies using the strategy of Warren et al. (2015) for rescaffolding a spruce (*Picea glauca*) genome. Here, we performed 15 iterations of LINKS using various settings. Fourth, we used the long read Pacific Biosciences data to further scaffold the assemblies using SSPACE LongRead v1-1 (Boetzer and Pirovano 2014). Fifth, we used LR_Gapcloser (Xu et al. 2019) with the long-read data and the de novo assembly to fill assembly gaps. Lastly, we used the Hi-C sequence data to further scaffold the de novo assembly and correct mis-assemblies using the 3D-DNA pipeline (Durand et al. 2016; Dudchenko et al. 2017). All commands and inputs during the various stages of assembly are documented on GitHub (github.com/jdmanthey/certhia_genomes1). The final assembly had an N50 of six scaffolds with L50 of 64.36 Mbp, with a total number of scaffolds = 8,651 (supplementary table S4, Supplementary Material online; contig N50 = 2,091, contig L50 = 143.07 kbp).

Genome Annotation

To annotate TEs and repetitive elements in the *C. americana* genome, we used a multistep process to identify de novo repeats and overrepresented sequences, manually curate repetitive elements, and mask the genome with these elements to create a TE and repetitive element summary file. First, we used RepeatModeler v1.0.11 (Smit and Hubley 2008) to identify repeats based on homology, structure, and repetitiveness in the de novo assembly. RepeatModeler utilizes multiple programs in its pipeline: RECON (Bao and Eddy 2002), RepeatScout (Price et al. 2005), and Tandem Repeats Finder (Benson 1999). We refined the RepeatModeler output by filtering matches to closely related sequences in the RepBase vertebrate database v. 24.03 (Jurka et al. 2005) and creating consensus sequences of novel repetitive elements.

First, we removed any RepeatModeler output sequences that were ≥98% identical to RepBase sequences as any matches in the de novo assembly would be of sufficient similarity to mask from the RepBase sequences. Next, we used BLAST and bedtools (Quinlan and Hall 2010) to extract putative matches to novel repeats from the de novo assembly. We used these extracted sequences to create consensus sequences for novel repetitive elements using the following steps: 1) MAFFT(Katoh and Standley 2013) alignment using Geneious (BioMatters Ltd), 2) 50% majority consensus sequences in Geneious, 3) trimming any ambiguous nucleotides on the ends of newly created consensus sequences. For any incomplete novel repetitive elements, where the ends of the

elements were not recovered in the new consensus sequences, we repeated the prior procedure to extract sequences from the reference genome with 1,000 bp flanks on each side of each BLAST match. We then repeated alignment and consensus sequence creation from these extracted sequences. This process was repeated up to three times as necessary.

We BLASTed all novel repeats against the RepBase database to assess any similarity via homology of our new sequences to previously characterized elements. Any similarity with previously characterized elements was used for naming purposes. Lastly, using the RepBase vertebrate database and all novel repeat elements, we used RepeatMasker v4.08 (Smit et al. 2015) to mask and summarize repetitive and TEs in the de novo *C. americana* genome.

We used the MAKER v2.31.10 pipeline (Cantarel et al. 2008) to annotate putative genes in the *C. americana* genome. First, we used MAKER to predict genes using proteins from other Passeriformes species: *Parus major* (GCF_001522545.3_Parus_major1.1_protein.faa), *Ficedula albicolis* (GCF_000247815.1_FicAlb1.5_protein.faa), and *T. guttata* (GCF_000151805.1_Taeniopygia_guttata-3.2.4_protein.faa) (Warren et al. 2010; Ellegren et al. 2012; Laine et al. 2016). After this first round of MAKER, we used the initial MAKER predictions to train SNAP (Korf 2004) and Augustus (Stanke and Waack 2003). Finally, using the SNAP- and Augustus-trained models, we ran a second iteration of MAKER to predict gene models in the *C. americana* genome. We then used BUSCO v3 (Simão et al. 2015) with the tetrapod single orthologous gene set (set: odb9) to assess genome assembly completeness.

Creeper-Specific Mutation Rate

We extracted the putative *C. americana* coding sequence (CDS) from the de novo assembly using the MAKER output and bedtools. We downloaded the CDS sequences for *Parus major*, *Fidecula albicola*, and *T. guttata* (same versions as proteins) for homology-based comparisons. We performed a reciprocal BLAST of all species versus *C. americana* using BlastN (Camacho et al. 2009) to identify putative homologues across data sets.

To put the evolution of all the CDS regions in a timed evolutionary context, we downloaded a Passeriformes family-scale phylogenetic tree (Oliveros et al. 2019) and pruned the tree to the four representative families covered by our CDS downloads and novel assembly using the R package ape (Paradis et al. 2004): Certhiidae, Estrildidae, Muscicapidae, and Paridae.

We used T-Coffee (Notredame et al. 2000) to align the putative homologues between the four passerine species. T-Coffee translates nucleotide sequences, aligns them using several alignment algorithms, takes the averaged best alignment of all alignments, and back translates the protein alignments to provide a nucleotide alignment for each gene. Prior to back-translating, we removed any gaps in the protein alignments using trimAl (Capella-Gutiérrez et al. 2009).

With the alignments for all genes, we tested for selection using the gene-wide and branch-specific tests for selection utilized in CODEML (Yang 1997). Any alignments with gene-wide or branch-specific evidence for selection were removed for mutation-rate analyses, after correcting for multiple tests using the Benjamini and Hochberg (1995) method to control false discovery rate. From each gene alignment, we used the R packages rphast, Biostrings, and seqinr (Charif and Lobry 2007; Hubisz et al. 2011; Pagès et al. 2017) to extract 4-fold degenerate sites from each alignment. We concatenated the 4-fold degenerate sites ($N \sim 1.4$ million) and used jModelTest (Darriba et al. 2012) to determine an appropriate model of sequence evolution. Lastly, we used the GTR + I model of sequence evolution in PhyML and a user-specified tree (from Oliveros et al. [2019]) to estimate branch lengths based on the 4-fold degenerate sites. Lastly, we used the *Certhia*-specific branch length of this tree along with divergence time estimates (also from Oliveros et al. [2019]) to estimate a mean and 95% HPD distribution of potential *Certhia*-lineage-specific mutation rates (2.506×10^{-9} substitutions/site/year; 95% HPD $= 2.243 \times 10^{-9}$ to 2.839×10^{-9}). This rate is generally in line with other Passeriformes species such as *Taeniopygia guttatta* (2.21×10^{-9} substitutions/site/year) (Nam et al. 2010), but slower than the relatively fast mutation rate of *Zosterops lateralis* (3.16×10^{-9} substitutions/site/year) (Cornetti et al. 2015).

Population Genomic Resequencing and Initial Data Processing and Filtering

For population genomic resequencing, we used 24 *C. americana* and one *C. familiaris* (outgroup), including *C. americana* from regions far from the putative contact zones as well as neighboring populations (fig. 2 and supplementary table S1, Supplementary Material online). We extracted genomic DNA using QIAGEN (Hilden, Germany) DNeasy blood and tissue kits following manufacturer guidelines. From genomic extractions, DNA was used to create Illumina sequencing libraries and sequenced on either an Illumina HiSeq3500 or NovaSeq6000 at the Oklahoma Medical Research Foundation (OMRF) Clinical Genomics Center. We aimed to sequence each individual at 10–25× genomic coverage, but several individuals failed to reach this expectation (supplementary fig. S1, Supplementary Material online).

We used bbduk, part of the bbmap package (Bushnell 2014), to trim adapters and quality filter raw sequencing data. We used the BWA-MEM implementation of the Burrows–Wheeler algorithm in BWA (Li and Durbin 2009) to align filtered reads to the de novo *C. americana* genome. We used samtools v1.4.1 (Li et al. 2009) to convert the BWA

output SAM file to BAM format, and lastly cleaned, sorted, added read groups to, and removed duplicates from each BAM file using the Genome Analysis Toolkit (GATK) v4.1.0.0 (McKenna et al. 2010). With the bam alignment files, we estimated depth of sequencing using the samtools "depth" command. Lastly, we used GATK's HaplotypeCaller to call initial genotypes for each individual and then used the GATK function GenotypeGVCFs to group genotype all individuals together, for both variant and invariant sites. We used VCFtools v0.1.14 (Danecek et al. 2011) to initially filter all variant and invariant site calls using the following restrictions: 1) minimum site quality of 20, 2) minimum genotype quality of 20, 3) minimum depth of coverage of 5, and 4) maximum mean depth of coverage of 50. Some analyses used additional restrictions on data quality and are detailed in the appropriate sections of the methods (supplementary table S2, Supplementary Material online).

Population Genomics

Recombination and LD

We used the LDhat software (McVean and Auton 2007) to estimate effective rates of recombination across the *C. americana* genome for both the northern and southern *Certhia* lineages. For use in LDhat, we filtered variants to only include biallelic SNPs with a maximum 35% missing proportion of individuals genotyped for a SNP to be included. We created a modified likelihood lookup table from the LDhat precomputed tables using a sample size of 12 (sampling for each lineage) and a population mutation rate parameter estimate of 0.001, that is, the closest value to the empirical *Certhia* value. We used the LDhat "interval" module to implement a Bayesian MCMC sampling algorithm to estimate effective recombination rates across each scaffold. We ran this module for five million iterations sampling once per 5,000. We used the LDhat module "stat" to summarize the output, discarding the first 20% of samples as burn-in, and summarized the LDhat output in 100 kbp windows for each of the two *Certhia* lineages. A recent simulation study investigating sample size effects on LDhat recombination rate estimates showed reliable recombination landscape inference with as few as ten haploid genomes (Stukenbrock and Dutheil 2018), suggesting our sample sizes are sufficient for reasonable recombination rate estimation.

We estimated LD in the R package "genetics" (Warnes et al. 2019). Here, LD was calculated as the squared correlation coefficient between markers. We calculated pairwise LD for SNPs on chromosomes $\geq$10 Mbp, with a MAF $\geq$0.2, and thinned to at least 50 kbp separation and a maximum of 1,000 SNPs per chromosome.

Diversity, Divergence, and Differentiation

For all estimates of genetic diversity and differentiation we only used biallelic variants with a maximum 40% missing individuals for a variant to be included. We used custom R scripts to measure genetic diversity (observed heterozygosity) within lineages and both relative and absolute genetic differentiation between lineages, F_{ST} and D_{XY}, respectively. We used the Reich et al. (2009) estimator of F_{ST} because this has been shown to be an unbiased F_{ST} estimator when using low sample sizes and high numbers of genetic markers (Willing et al. 2012). For estimates of H_O, F_{ST}, and D_{XY}, we only used SNPs. We additionally estimated pairwise estimates of F_{ST} and D_{XY} between all sampling locations to obtain genomic mean values (supplementary table S5, Supplementary Material online). Because of the small sample sizes per site, we refrained from using window-based F_{ST} and D_{XY} estimates in these pairwise comparisons. Additionally, using both SNPs and indels, we estimated the number of fixed differences between lineages, and summarized these metrics in 100 kbp windows.

Because some samples had lower relative sequencing coverage, we investigated the relationship between sequencing coverage and genomic heterozygosity per individual. Overall, we found no relationship between the two (supplementary fig. S10, Supplementary Material online), suggesting our variant filtering helps preclude diversity estimate biases from sites with low coverage sequencing.

To assess whether a MAF filter impacted estimates of F_{ST} (Linck and Battey 2019), we estimated F_{ST} in windows with a MAF = 0.05 and compared estimates with the full data set lacking a MAF filter. Here, F_{ST} results were largely unaffected by the MAF filter, with a strong positive correlation between data sets ($r = 0.999$, slope = 1.062; supplementary fig. S11, Supplementary Material online).

To look at patterns of segregating sites in allopatric versus sky island populations, we first found SNPs that were fixed differences between the allopatric Utah and Mexico populations. Using these SNPs as putative segregating sites between lineages, we calculated the proportion of northern and southern alleles at these SNPs for each of the sky island populations.

Genomic Window Correlations

We used the R package Hmisc (Harrell and Dupont 2020) to estimate sliding window correlations between genic content, TE content, recombination rates, genetic diversity, and genetic divergence and differentiation. Here, we used three data sets to estimate correlations: 1) all windows, 2) windows thinned to 5% (one every 20), and 3) windows thinned to 2% (one every 50). We thinned windows to estimate correlations to identify whether correlations persist when sampling number decreases and the effects of linkage are reduced. We estimated correlation coefficients using both Spearman and Pearson correlation coefficients, and assess significance with

an $\alpha \leq 0.05$ following a Bonferroni correction (Bonferroni 1936). All correlations are presented in supplementary table S3, Supplementary Material online.

Because we identified correlations between population genomic statistics and chromosomal characteristics at both the chromosomal (fig. 3) and sliding windows (fig. 4*A*) levels, we wanted to parse out if there were any chromosomal effects not accounted for by the gene content, TE content, and local recombination rate in sliding windows. Here, we used partial regression to extract the residuals when genomic differentiation, divergence, or diversity are the response variables estimated by the explanatory variables gene content, TE content, and recombination rate (e.g., the R syntax: "resid(lm(fst ~ genes + repeats + rho))"). Then we calculated the mean residual per chromosome (mean across all windows) and regressed that with log chromosome size.

Demography

We used fastSimcoal2 v2.6.0.3 (Excoffier et al. 2013) to assess the fit of multiple demographic scenarios to our data set: 1) pure isolation, 2) isolation with migration, 3) speciation with gene flow, and 4) secondary contact (supplementary fig. S4, Supplementary Material online). As input, fastSimcoal2 uses a two-population SFS file, which we generated using the program easysfs (github.com/isaacovercast/easySFS). Here, we used sites genotyped in all individuals (i.e., no missing data; supplementary table S2, Supplementary Material online). We implemented 100 replicates of fastSimcoal2 to fit the SFS to each demographic scenario. For each estimated parameter, we used wide search ranges with uniform or log-uniform prior distributions (available at: https://github.com/jdmanthey/certhia_genomes1/tree/master/16_revisions/07_fastsimcoal). For each replicate, we used 200,000 coalescent simulations and 20 expectation–maximization cycles. For each demographic model, we extracted the estimated parameters from the replicate that maximized the likelihood. We then ran fastSimcoal2 100 times with the parameters that maximize the likelihood for each model to obtain simulated SFS and associated likelihood distributions for each scenario, thereby informing us about variance in likelihood estimations for each model in fastSimcoal2 (Meier et al. 2017). For the best model (isolation with secondary contact), the estimated likelihood from the empirical SFS (−951,722) was both: 1) slightly higher than the range of best likelihoods from the simulated SFS runs (range = −951,977 to −951,738; supplementary fig. S5, Supplementary Material online) and 2) close to the estimated maximum observable likelihood given the empirical SFS (−942,330). We used the approach of Meier et al. (2017) to estimates confidence intervals for the best fitting model with a nonparametric block-bootstrap approach. Here, we created 100 data sets by sampling (with replacement) 500 SNP blocks, converting these data sets to the SFS, and running 20 replicates of fastSimcoal2 for each of the 100 data sets using the same settings as with the original data set. We took the parameter estimates from the best model for each of the 100 data sets to estimate confidence intervals for each parameter estimate. We included our empirical estimate of mutation rate in fastSimcoal2 to allow us to estimate timing in absolute numbers. Because of limited information about *Certhia* generation times, we used two possible generation times: 1) double and 2) triple the age of sexual maturity (i.e., 2 or 3 years with sexual maturity at 1 year).

We also used the program smc++ v1.15.2 (Terhorst et al. 2017) to estimate effective population size changes through time for each of the *Certhia* lineages, including the lineage splitting time. Because the best demographic model identified with fastSimcoal2 included gene flow after secondary contact, the smc++ model will be somewhat biased because it does not account for gene flow between lineages. However, our main goal with using smc++ was to obtain potential fluctuations in effective population sizes through time that could be used as a starting point to account for demographic histories in diploS/HIC (Kern and Schrider 2018) tests for selection. smc++ uses information about the spatial arrangement of unphased SNPs along chromosomes to infer variation in N_E through time, utilizing information from variants' SFS and LD patterns. We limited analyses to individuals with greater than 10× mean genomic coverage to reduce inclusion of individuals with large portions of the genome ungenotyped (northern lineage $N = 12$ individuals; southern lineage $N = 7$ individuals).

Introgression

We used D and f_d statistics to identify signatures of introgression between the two *Certhia* lineages (Green et al. 2010; Durand et al. 2011; Martin et al. 2015). Both the D and f_d test statistics use "ABBA–BABA" relationships between three ingroup populations plus an outgroup with the assumed evolutionary relationship: (((P1, P2), P3), O). In the ABBA–BABA tests, the "A" and "B" represent the ancestral and derived alleles, respectively. With neutral lineage sorting, we would expect equal frequencies of ABBA and BABA patterns among taxa, with significant deviations indicative of introgression. Here, we tested introgression into any of the Arizona sky island populations from parental populations using six phylogenetic topologies (fig. 6). We calculated the introgression statistics for each of the comparisons in 100 kbp sliding windows, limited to windows with ≥500 SNPs. We decided to calculate both of these introgression statistics because D has been shown to have some biases in genomic regions with low genetic diversity (Martin et al. 2015), for which the f_d statistic is not particularly biased. Indeed, some genomic regions that indicate introgression with the D statistic do not match trends with the f_d statistic (supplementary fig. S12, Supplementary Material online) and may have been misinterpreted if only the D statistic was used.

Population Genetic Structure

We used a subset of our SNPs that were genotyped in each individual and were a minimum of 10 kbp apart along scaffolds to assess population genetic structure. First, we used the program STRUCTURE (Pritchard et al. 2000) to determine ancestry of each individual to either the northern or southern lineage. We initially ran STRUCTURE to infer the lambda parameter while estimating the likelihood of one population ($k = 1$) (Pritchard et al. 2000). We then used the inferred value of lambda for subsequent analyses, where we performed ten replicates of likelihood estimation for two genetic clusters. We assumed correlated allele frequencies, an admixture model, and performed analyses for a burn-in of 50,000 steps and a subsequent 50,000 MCMC iterations. We also estimated genetic structure between the lineages using DAPC (Jombart et al. 2010), implemented in the adegenet R package (Jombart and Ahmed 2011). DAPC implements principal components analysis of all genetic variants followed by discriminant analysis to determine appropriate genetic clusters of individuals. We ran the "dapc" function for 100,000 iterations to determine groupings, chose the appropriate number of genetic clusters with BIC (supplementary fig. S13, Supplementary Material online), and visualized results with the "compoplot" function.

We also assessed whether STRUCTURE results were consistent with different thinning of SNPs. To do this, we used a subset of our SNPs that were genotyped in each individual and were a minimum of 50 kbp apart along scaffolds. Results were consistent between this 50 kbp thinned data set (supplementary fig. S14, Supplementary Material online) and the 10 kbp thinned data set (fig. 2).

Evolutionary Relationships

We inferred population relationships using TreeMix v1.13 (Pickrell and Pritchard 2012). TreeMix uses SNPs to infer a maximum likelihood population or species tree, and subsequently adds migration edges to populations that are more closely related than can be explained by the bifurcating topology alone. We ran TreeMix with biallelic SNPs present in all individuals, and a minimum of 10 kbp separation between SNPs to help reduce effects of linkage. We added migration edges to the phylogeny until they explained less than 0.02% of the variance in the SNP data (Pickrell and Pritchard 2012). Lastly, we assessed support for this population tree by performing 100 bootstraps of the analysis, using 200 SNP sampling blocks for bootstraps.

We also estimated "gene trees" for nonoverlapping 50 kbp sliding windows using RAxML v8.2.12 (Stamatakis 2014) with the GTRGAMMA model of sequence evolution. For a site (i.e., single bp) to be included, we required it to be genotyped in at least 60% ($n = 15$) of individuals. For a window to be included, we required: 1) a minimum of 50% of sites in the alignment (i.e., 25 kbp) and 2) no individuals with less than 10% (5 kbp) of the alignment genotyped. We used 100 rapid bootstraps to estimate support for each tree and determine a "best-supported" tree for each window. We summarized the best-supported phylogenies ($n = 18{,}470$) using the sumtrees.py script, part of the DendroPy Python package (Sukumaran and Holder 2010) (supplementary fig. S15, Supplementary Material online). Additionally, we used ASTRAL III v 5.6.3 (Zhang et al. 2018) to calculate a species tree from all of the sliding windows' phylogenies (supplementary fig. S16, Supplementary Material online). For the ASTRAL analysis, we used the quartet frequencies as a measure of local support (Sayyari and Mirarab 2016).

Natural Selection

We used the machine-learning program diploS/HIC (Kern and Schrider 2018) to predict hard and soft selective sweeps and variation linked to selective sweeps. diploS/HIC uses a deep convolutional neural network to identify sweeps in sliding windows along the genome with population genomic data (Kern and Schrider 2018). First, we performed 2,000 coalescent simulations using discoal (Kern and Schrider 2016) to simulate multiple scenarios for each lineage: hard selective sweeps, soft selective sweeps, variation linked to those two categories of selective sweeps, and neutral evolution. We used 220 kbp windows divided into 11 subwindows. The simulations were performed using the smc++ estimated demographic histories.

Demographic and mutation rate uncertainty were incorporated into the theta prior, allowing the contemporary effective population size to vary between 1/3 to 3× the smc++ estimates. We used a truncated exponential prior on recombination rate per bp that encompassed a majority of values estimated from our data (fig. 4*A*): (0.0033, 0.0754). We used a uniform prior on selection coefficients $\sim$ (0.00025, 0.025), and conditioned sweep completion occurring between 10,000 generations ago and the present. We used uniform priors on adaptive variant initial frequencies for soft sweeps as $\sim$ (0, 0.2).

We used these simulations as input for training a supervised machine-learning algorithm to differentiate between neutral evolution, sweeps, and windows linked to selection. The training incorporated a genome-wide sequence mask to incorporate the empirical missing data features with the simulation data. After training, we classified genomic regions for each of the *Certhia* lineages using two scales: 20- and 50 kbp windows.

Investigation of Putative Inversions

We identified three regions of the genome greater than 1 Mbp that showed a strong signature of introgression as well as reduced recombination rates. We hypothesized that these may be ancestrally segregating inversions and aimed to

use multiple lines of evidence to assess this hypothesis. First, we extracted SNPs from these regions with no missing data to estimate genetic structure using the program STRUCTURE. Second, we compared the patterns of recombination rate and LD in these regions relative to the overall recombination rates and LD for the chromosomes these regions are located in. Third, we estimated overall heterozygosity of these regions relative to the chromosomal background. Lastly, we measured per individual heterozygosity in the putative inversions in SNPs segregating between *Certhia* lineages (MAF $\geq$0.3).

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

We would like to thank the collection managers, curators, and contributors for generous tissue loans for all individuals used in this study from the following museums: Museum of Southwestern Biology, Burke Museum of Natural History and Culture, Denver Museum of Nature and Science, and University of Kansas Biodiversity Institute. Sequencing was supported by Texas Tech University start-up funding to J.D.M. and a generous donation from the Ferguson family to G.M.S. and the Denver Museum of Nature and Science. The High-Performance Computing Center at TTU supported computational analyses.

Author Contributions

All authors designed the study and conducted field work to acquire samples. J.D.M. performed laboratory and bioinformatic work and wrote the first draft of the manuscript. All authors contributed to revising and improving the final draft of the manuscript.

Data Availability

The genome assembly and raw reads for the genome assembly are available through NCBI BioProject PRJNA604561. All raw resequencing data are available through NCBI BioProject PRJNA605140. All codes are available on GitHub (github.com/jdmanthey/certhia_genomes1).

Literature Cited

Adrion JR, Galloway JG, Kern AD. 2020. Predicting the landscape of recombination using deep learning. Mol Biol Evol. 37(6):1790–1808.

Bao Z, Eddy SR. 2002. Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12(8):1269–1276.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B Methodol. 57(1):289–300.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Boetzer M, Pirovano W. 2014. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 15:211.

Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilita. Libreria internazionale Seeber.

Branca A, et al. 2011. Whole-genome nucleotide diversity, recombination, and linkage disequilibrium in the model legume *Medicago truncatula*. Proc Natl Acad Sci U S A. 108(42):E864–E870.

Brand CM, White FJ, Ting N, Webster TH. 2021. Soft sweeps predominate recent positive selection in bonobos (*Pan paniscus*) and chimpanzees (*Pan troglodytes*). bioRxiv. doi: 10.1101/2020.12.14.422788.

Burri R. 2017. Interpreting differentiation landscapes in the light of long-term linked selection. Evol Lett. 1(3):118–131.

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(11):1656–1665.

Bushnell B. 2014. BBMap: a fast, accurate, splice-aware aligner. Berkeley (CA): Ernest Orlando Lawrence Berkeley National Laboratory.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10(1):421.

Cantarel BL, et al. 2008. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1):188–196.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25(15):1972–1973.

Carneiro M, et al. 2014. The genomic architecture of population divergence between subspecies of the European rabbit. PLoS Genet. 10(8):e1003519.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: Bastolla U, Porto M, Roman H, Vendruscolo M, editors. Structural approaches to sequence evolution. New York: Springer. p. 207–232.

Cornetti L, et al. 2015. The genome of the "great speciator" provides insights into bird diversification. Genome Biol Evol. 7(9):2680–2691.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol Ecol. 23(13):3133–3157.

Cutter AD, Payseur BA. 2003. Selection at linked sites in the partial selfer *Caenorhabditis elegans*. Mol Biol Evol. 20(5):665–673.

Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet. 14(4):262–274.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27(15):2156–2158.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 9(8):772.

Davey JW, et al. 2017. No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. Evol Lett. 1(3):138–154.

Delmore KE, et al. 2015. Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. Mol Ecol. 24(8):1873–1888.

Delmore KE, et al. 2018. Comparative analysis examining patterns of genomic differentiation across multiple episodes of population divergence in birds. Evol Lett. 2(2):76–87.

Dobzhansky T. 1937. Genetic nature of species differences. Am Nat. 71:404–420.

Dudchenko O, et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356(6333):92–95.
Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol Biol Evol. 28(8):2239–2252.
Durand NC, et al. 2016. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3(1):99–101.
Dutoit L, Burri R, Nater A, Mugal CF, Ellegren H. 2017. Genomic distribution and estimation of nucleotide diversity in natural populations: perspectives from the collared flycatcher (*Ficedula albicollis*) genome. Mol Ecol Resour. 17(4):586–597.
Dutoit L, Vijay N, et al. 2017. Covariation in levels of nucleotide diversity in homologous regions of the avian genome long after completion of lineage sorting. Proc R Soc B. 284(1849):20162756.
Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. Trends Ecol Evol. 25(5):283–291.
Ellegren H, et al. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature 491(7426):756–760.
Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet. 9(10):e1003905.
Faria R, Johannesson K, Butlin RK, Westram AM. 2019. Evolving inversions. Trends Ecol Evol. 34(3):239–248.
Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. Trends Genet. 28(7):342–350.
Futuyma DJ, Mayer GC. 1980. Non-allopatric speciation in animals. Syst Biol. 29(3):254–271.
Gagnaire PA, Pavey SA, Normandeau E, Bernatchez L. 2013. The genetic architecture of reproductive isolation during speciation-with-gene-flow in lake whitefish species pairs assessed by RAD sequencing. Evolution 67(9):2483–2497.
Green RE, et al. 2010. A draft sequence of the Neandertal genome. Science 328(5979):710–722.
Guerrero RF, Hahn MW. 2017. Speciation as a sieve for ancestral polymorphism. Mol Ecol. 26(20):5362–5368.
Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol Ecol. 27(11):2477–2497.
Han F, et al. 2017. Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. Genome Res. 27(6):1004–1015.
Harrell FE, Dupont C. 2020. Hmisc: Harrell miscellaneous. Version 4.3-1. R: R. Available from: http://biostat.mc.vanderbilt.edu/Hmisc.
Henderson EC, Brelsford A. 2020. Genomic differentiation across the speciation continuum in three hummingbird species pairs. BMC Evol Biol. 20(1):113–111.
Hooper DM, Griffith SC, Price TD. 2019. Sex chromosome inversions enforce reproductive isolation across an avian hybrid zone. Mol Ecol. 28(6):1246–1262.
Hooper DM, Price TD. 2017. Chromosomal inversion differences correlate with range overlap in passerine birds. Nat Ecol Evol. 1(10):1526–1534.
Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. Brief Bioinform. 12(1):41–51.
Irwin DE, Alcaide M, Delmore KE, Irwin JH, Owens GL. 2016. Recurrent selection explains parallel evolution of genomic regions of high relative but low absolute differentiation in a ring species. Mol Ecol. 25(18):4488–4507.
Janoušek V, Munclinger P, Wang L, Teeter KC, Tucker PK. 2015. Functional organization of the genome may shape the species boundary in the house mouse. Mol Biol Evol. 32(5):1208–1220.
Jensen-Seaman MI, et al. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res. 14(4):528–538.
Jombart T, Ahmed I. 2011. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics 27(21):3070–3071.
Jombart T, Devillard S, Balloux F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 11:94.
Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110(1–4):462–467.
Kaback DB, Guacci V, Barber D, Mahon JW. 1992. Chromosome size-dependent control of meiotic recombination. Science 256(5054):228–232.
Kapusta A, Suh A. 2017. Evolution of bird genomes—a transposon's-eye view. Ann N Y Acad Sci. 1389(1):164–185.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.
Kawakami T, et al. 2014. A high-density linkage map enables a second-generation collared flycatcher genome assembly and reveals the patterns of avian recombination rate variation and chromosomal evolution. Mol Ecol. 23(16):4035–4058.
Kern AD, Schrider DR. 2016. Discoal: flexible coalescent simulations with selection. Bioinformatics 32(24):3839–3841.
Kern AD, Schrider DR. 2018. diploS/HIC: an updated approach to classifying selective sweeps. G3 (Bethesda) 8(6):1959–1970.
Koonin EV. 2009. Evolution of genome architecture. Int J Biochem Cell Biol. 41(2):298–306.
Koren S, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 27(5):722–736.
Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.
Kraft T, Säll T, Magnusson-Rading I, Nilsson N-O, Halldén C. 1998. Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). Genetics 150(3):1239–1244.
Kulathinal RJ, Bennett SM, Fitzpatrick CL, Noor MA. 2008. Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. Proc Natl Acad Sci U S A. 105(29):10051–10056.
Laine VN, et al. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. Nat Commun. 7:10474.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25(14):1754–1760.
Li H, et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079.
Linck E, Battey C. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. Mol Ecol Resour. 19(3):639–647.
Lomolino MV, Davis R. 1997. Biogeographic scale and biodiversity of mountain forest mammals of western North America. Glob Ecol Biogeogr Lett. 6(1):57–76.
Lomolino MV, Riddle BR, Brown JH. 2006. Biogeography. Sunderland (MA): Sinauer Associates, Inc.
Lotterhos KE. 2019. The effect of neutral recombination variation on genome scans for selection. G3 (Bethesda) 9(6):1851–1867.
Lundberg M, et al. 2017. Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks. Evol Lett. 1(3):155–168.
Lynch M, Walsh B. 2007. The origins of genome architecture. Sunderland (MA): Sinauer Associates.
Manthey JD, Klicka J, Spellman GM. 2011a. Cryptic diversity in a widespread North American songbird: phylogeography of the Brown Creeper (*Certhia americana*). Mol Phylogenet Evol. 58(3):502–512.
Manthey JD, Klicka J, Spellman GM. 2011b. Isolation-driven divergence: speciation in a widespread North American songbird (Aves: Certhiidae). Mol Ecol. 20(20):4371–4384.

Manthey JD, Klicka J, Spellman GM. 2014. Effects of climate change on the evolution of Brown Creeper (*Certhia americana*) lineages. Auk Ornitholog Adv. 131(4):559–570.
Manthey JD, Robbins MB, Moyle RG. 2016. A genomic investigation of the putative contact zone between divergent Brown Creeper (*Certhia americana*) lineages: chromosomal patterns of genetic differentiation. Genome 59(2):115–125.
Marshall JT. 1956. Summer birds of the Rincon Mountains, Saguaro National Monument, Arizona. Condor. 58:81–97.
Martin SH, Davey JW, Jiggins CD. 2015. Evaluating the use of ABBA–BABA statistics to locate introgressed loci. Mol Biol Evol. 32(1):244–257.
Martin SH, Davey JW, Salazar C, Jiggins CD. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol. 17(2):e2006288.
Matthey-Doret R, Whitlock MC. 2019. Background selection and F_{ST}: consequences for detecting local adaptation. Mol Ecol. 28(17):3902–3914.
Mayr E. 1942. Systematics and the origin of species, from the viewpoint of a zoologist. Cambridge (MA): Harvard University Press.
McKenna A, et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297–1303.
McVean G, Auton A. 2007. LDhat 2.1: a package for the population genetic analysis of recombination. Oxford: Department of Statistics.
Meier JI, et al. 2017. Demographic modelling with whole-genome data reveals parallel origin of similar Pundamilia cichlid species after hybridization. Mol Ecol. 26(1):123–141.
Nachman MW, Payseur BA. 2012. Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. Philos Trans R Soc Lond B Biol Sci. 367(1587):409–421.
Nam K, et al. 2010. Molecular evolution of genes in avian genomes. Genome Biol. 11(6):R68.
Nosil P, Feder JL. 2012. Genomic divergence during speciation: causes and consequences. Philos Trans R Soc Lond B Biol Sci. 367(1587):332–342.
Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302(1):205–217.
Oliveros CH, et al. 2019. Earth history and the passerine superradiation. Proc Natl Acad Sci U S A. 116(16):7916–7925.
Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2017. Biostrings: efficient manipulation of biological strings. R Package Version 2.0.
Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20(2):289–290.
Payseur BA, Rieseberg LH. 2016. A genomic perspective on hybridization and speciation. Mol Ecol. 25(11):2337–2360.
Phung TN, Huber CD, Lohmueller KE. 2016. Determining the effect of natural selection on linked neutral divergence across species. PLoS Genet. 12(8):e1006199.
Pickrell JK, Pritchard JK. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8(11):e1002967.
Poelstra JW, et al. 2014. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science 344(6190):1410–1414.
Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1):i351–i358.
Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. J Hered. 155(2):945–959.
Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.
Ravinet M, et al. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. J Evol Biol. 30(8):1450–1477.
Reich D, Thangaraj K, Patterson N, Price AL, Singh L. 2009. Reconstructing Indian population history. Nature 461(7263):489–494.
Renaut S, et al. 2013. Genomic islands of divergence are not affected by geography of speciation in sunflowers. Nat Commun. 4:1827.
Riesch R, et al. 2017. Transitions between phases of genomic differentiation during stick-insect speciation. Nat Ecol Evol. 1(4):82.
Rockman MV. 2012. The QTN program and the alleles that matter for evolution: all that's gold does not glitter. Evolution 66(1):1–17.
Roesti M, Moser D, Berner D. 2013. Recombination in the threespine stickleback genome—patterns and consequences. Mol Ecol. 22(11):3014–3027.
Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7):1654–1668.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.
Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013–2015. Institute for Systems Biology. Available from: http://repeatmasker. org.
Smit AF, Hubley R. 2008. RepeatModeler Open-1.0. Available from: http://www.repeatmasker.org.
Spencer CC, et al. 2006. The influence of recombination on human genetic diversity. PLoS Genet. 2(9):e148.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9):1312–1313.
Stanke M, Waack S. 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 19(Suppl 2):ii215–ii225.
Stankowski S, et al. 2019. Widespread selection and gene flow shape the genomic landscape during a radiation of monkeyflowers. PLoS Biol. 17(7):e3000391.
Stapley J, Feulner PG, Johnston SE, Santure AW, Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc B. 372(1736):20160455.
Stukenbrock EH, Dutheil JY. 2018. Fine-scale recombination maps of fungal plant pathogens reveal dynamic recombination landscapes and intragenic hotspots. Genetics 208(3):1209–1229.
Sukumaran J, Holder MT. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26(12):1569–1571.
Takahashi A, Liu Y-H, Saitou N. 2004. Genetic variation versus recombination rate in a structured population of mice. Mol Biol Evol. 21(2):404–409.
Terhorst J, Kamm JA, Song YS. 2017. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet. 49(2):303–309.
Tigano A, et al. 2021. Chromosome size affects sequence divergence between species through the interplay of recombination and selection. bioRxiv. doi: 10.1101/2021.01.15.426870.
Toews DP, et al. 2016. Plumage genes and little else distinguish the genomes of hybridizing warblers. Curr Biol. 26(17):2313–2318.
Van Doren BM, et al. 2017. Correlated patterns of genetic diversity and differentiation across an avian family. Mol Ecol. 26(15):3982–3997.
Vijay N, et al. 2016. Evolution of heterogeneous genome differentiation across multiple contact zones in a crow species complex. Nat Commun. 7(1):1–10.
Vijay N, et al. 2017. Genomewide patterns of variation in genetic diversity are shared among populations, species and higher-order taxa. Mol Ecol. 26(16):4284–4295.
Wade TG, Riitters KH, Wickham JD, Jones KB. 2003. Distribution and causes of global forest fragmentation. Conserv Ecol. 7(2):7.
Warnes G, Gorjanc G, Leisch F, Man M. 2019. Genetics: population genetics. Version 1.3.8.1.2.
Warren RL, et al. 2015. LINKS: scalable, alignment-free scaffolding of draft genomes with long reads. Gigascience 4:35.

Warren WC, et al. 2010. The genome of a songbird. Nature 464(7289):757–762.
Westram AM, et al. 2018. Clines on the seashore: the genomic architecture underlying rapid divergence in the face of gene flow. Evol Lett. 2(4):297–309.
Willing E-M, Dreyer C, Van Oosterhout C. 2012. Estimates of genetic differentiation measured by F_{ST} do not necessarily require large sample sizes when using many SNP markers. PLoS One 7(8):e42649.
Xu G-C, et al. 2019. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. Gigascience 8(1):giy157.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci. 13(5):555–556.
Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19(Suppl 6):153.

Associate editor: Kirk Lohmueller