

## Phylogenomic Data Reveal Widespread Introgression Across the Range of an Alpine and Arctic Specialist

ERIK R. FUNK<sup>1,\*</sup>, GARTH M. SPELLMAN<sup>2</sup>, KEVIN WINKER<sup>3</sup>, JACK J. WITHROW<sup>3</sup>, KRISTEN C. RUEGG<sup>4</sup>,  
ERIKA ZAVALA<sup>5</sup> AND SCOTT A. TAYLOR<sup>1</sup>

<sup>1</sup>Department of Ecology and Evolutionary Biology, University of Colorado Boulder, 1900 Pleasant St. 334 UCB, Boulder, CO 80309, USA; <sup>2</sup>Department of Zoology, Denver Museum of Nature and Science, 2001 Colorado Blvd., Denver, CO 80205, USA; <sup>3</sup>University of Alaska Museum, University of Alaska Fairbanks, 1962 Yukon Dr., Fairbanks, AK 99775, USA; <sup>4</sup>Department of Biology, Colorado State University, 251 W Pitkin St., Fort Collins, CO 80521, USA; and <sup>5</sup>Department of Ecology and Evolutionary Biology, University of California Santa Cruz, 1156 High St., Santa Cruz CA, 95064, USA

\*Correspondence to be sent to: Department of Ecology and Evolutionary Biology, University of Colorado Boulder, 1900 Pleasant St. 334 UCB, Boulder, CO 80309, USA;

E-mail: erik.funk@colorado.edu

Received 16 October 2019; reviews returned 28 August 2020; accepted 31 August 2020

Associate Editor: Stephen Smith

**Abstract.**—Understanding how gene flow affects population divergence and speciation remains challenging. Differentiating one evolutionary process from another can be difficult because multiple processes can produce similar patterns, and more than one process can occur simultaneously. Although simple population models produce predictable results, how these processes balance in taxa with patchy distributions and complicated natural histories is less certain. These types of populations might be highly connected through migration (gene flow), but can experience stronger effects of genetic drift and inbreeding, or localized selection. Although different signals can be difficult to separate, the application of high-throughput sequence data can provide the resolution necessary to distinguish many of these processes. We present whole-genome sequence data for an avian species group with an alpine and arctic tundra distribution to examine the role that different population genetic processes have played in their evolutionary history. Rosy-finches inhabit high elevation mountaintop sky islands and high-latitude island and continental tundra. They exhibit extensive plumage variation coupled with low levels of genetic variation. Additionally, the number of species within the complex is debated, making them excellent for studying the forces involved in the process of diversification, as well as an important species group in which to investigate species boundaries. Total genomic variation suggests a broadly continuous pattern of allele frequency changes across the mainland taxa of this group in North America. However, phylogenomic analyses recover multiple distinct, well supported, groups that coincide with previously described morphological variation and current species-level taxonomy. Tests of introgression using D-statistics and approximate Bayesian computation reveal significant levels of introgression between multiple North American taxa. These results provide insight into the balance between divergent and homogenizing population genetic processes and highlight remaining challenges in interpreting conflict between different types of analytical approaches with whole-genome sequence data. [ABBA-BABA; approximate Bayesian computation; gene flow; phylogenomics; speciation; whole-genome sequencing.]

Population divergence and speciation underlie the generation of biodiversity. However, the genetic and environmental conditions under which these processes can occur remain an active area of research (Cruickshank and Hahn 2014; Morales et al. 2017; Wolf and Ellegren 2017; Beckman et al. 2018; Marques et al. 2019). Understanding population genetic structure and connectivity is central to determining how different processes have shaped evolutionary history and for evaluating the conditions under which divergence and subsequent speciation occur. Geographic isolation and selection can result in population differentiation, possibly leading to local adaptation, whereas gene flow reduces differentiation (Mayr 1942; Slatkin 1987; Caramunt et al. 2012; Gaither et al. 2015). Although the patterns produced by genetic drift, selection, and gene flow can be fairly predictable when considered independently, population genetic processes are not mutually exclusive. A population might be affected by processes that result in similar patterns (e.g. selection and drift; Miller and Lambert 2004; Eimes et al. 2011; Sutton et al. 2011), or opposing patterns (e.g. divergent selection and gene flow; Pinho and Hey 2010; Ellegren et al. 2012; Campbell et al. 2018). When this occurs, it

is more difficult to estimate the relative effects of each process on population differentiation.

Understanding the balance between evolutionary processes that result in divergence and homogenization of lineages is critical to our ability to understand evolutionary history. However, distinguishing between evolutionary processes can be particularly difficult for nonmodel organisms that are unevenly distributed (Keyghobadi et al. 2005), or characterized by complicated natural and evolutionary histories such as a metapopulation structure, irruptive population sizes, or elevational migration (Richards 2000; James et al. 2015). These types of distributional and demographic characteristics impact effective population size and migration (gene flow) between populations, altering the conditions under which divergence might occur. These complexities are likely to vary in nonmodel systems, highlighting the importance of understanding the balance between different population genetic processes in divergence and speciation. Small, isolated populations, such as those found on islands or mountaintops, might be particularly susceptible to stochastic effects such as inbreeding and genetic drift, increasing the degree of genetic divergence among

populations. This may be especially true for organisms with limited dispersal capabilities, or that face large geographic barriers, such as orthopterans across river basins (González-Serna et al. 2018), or amphibians across mountain ranges (Sánchez-Montes et al. 2018). The effects from inbreeding and genetic drift could put small populations at an increased risk of extirpation or extinction. However, high connectivity among isolated populations, as potentially seen in organisms that are nomadic or irruptive, is likely to homogenize populations and decrease divergence (e.g. redpoll finches; Mason and Taylor 2015). Increased connectivity might also alleviate risk of extirpation, and organisms with high dispersal abilities, such as birds, may be able to link fragmented or patchy distributions (Teitelbaum and Mueller 2019). Because isolation and connectivity have direct implications for the survival of small populations, understanding evolutionary history and population genetic processes can be a powerful tool for both describing and conserving biodiversity (Schwartz et al. 2007).

Studies of how genetic drift and gene flow interact to affect population differentiation have been conducted for multiple species. Cameron et al. (2019) examined how variation in both genetic drift (effective population size), and migration rates relate to patterns of population differentiation ( $F_{ST}$ ) in the Eastern red-backed salamander (*Plethodon cinereus*). Measures of population parameters across both fragmented and continuous habitats demonstrated that effective population size had a much stronger correlation with  $F_{ST}$  values than migration rates in patchy landscapes, whereas drift and migration explained  $F_{ST}$  patterns equally in continuous landscapes. Results from this study demonstrate that when distributions are patchy, or become fragmented, differentiation driven by drift can outweigh homogenization driven by gene flow. Llorens et al. (2017) found that even for wind-pollinated plants with high dispersal capabilities (*Allocaeusuarina humilis*), drift played a dominant role in driving differentiation in patchy population distributions when compared with the largely homogenized populations of more continuously distributed populations. Broadly, the results from these studies highlight that the increased magnitude of drift from patchy population distributions can drive differentiation despite the homogenizing effects of gene flow.

Although discerning different population demographic processes can be difficult, the high resolution offered by phylogenomic data is providing unprecedented levels of detail to examine hypotheses of evolutionary history (Green et al. 2010; Nadeau et al. 2013; Starrett et al. 2017). Here, we use whole-genome sequences to examine broad-scale population genomic and phylogenomic patterns in a closely related group of alpine and arctic specialist songbirds, the rosy-finches.

Rosy-finches are mid-sized (22–60 g) songbirds restricted to high elevations and cold environments, nesting in places upward of 4100 m (Lincoln 1916).

Their specialization to alpine sky islands and northern latitude arctic tundra produces a patchy distribution of breeding populations across their range (Fig. 1a). During winter, most leave their breeding grounds, undergoing short-distance movements and forming flocks at lower elevations and latitudes. They also exhibit a variety of morphological differences across this range (Fig. 1b), including variation in both size and plumage coloration (Macdougall-Shackleton et al. 2000). Although this group has a long history of taxonomic uncertainty (Macdougall-Shackleton et al. 2000), variation in morphology in North American taxa, as described by Johnson (1972), forms the basis of their most typical classification as four species (American Ornithologists' Union 1998), including the gray-crowned rosy-finch (*Leucosticte tephrocotis*), the brown-capped rosy-finch (*Leucosticte australis*), the black rosy-finch (*Leucosticte atrata*), and the Asian rosy-finch (*Leucosticte arctoa*). Despite distinct morphological differences (Fig. 1b), rosy-finches have been subjected to a suite of previous classifications, spanning a range from one to five species (American Ornithologists' Union 1931, 1983, 1998; Vaurie 1956; Howell et al. 1968; Drovetski et al. 2009; Dickinson and Christidis 2014; Clements et al. 2019).

An alternative to the four-species hypothesis described above is a two-species hypothesis that includes an Asian species and a single North American species, grouping the gray-crowned, black, and brown-capped rosy-finch taxa together (Dickinson and Christidis 2014). This treatment is supported by previous genetic data from a study including two nuclear genes (*ACO119*, *MC1R*) and one mitochondrial gene (*ND2*; Drovetski et al. 2009). Although these loci were found to lack population genetic structure across North America, it remains uncertain if this pattern is consistent across the rest of the genome, and a number of evolutionary scenarios remain plausible. The first is that these taxa are highly connected, as suggested by Drovetski et al. (2009), and ongoing gene flow has resulted in a pattern of widespread similarity across the genome. An alternative is that higher-resolution genetic data might reveal parts of the genome that do exhibit differentiation (e.g. Toews et al. 2016b). Although rosy-finches undertake seasonal movements during winter (e.g. *L. t. littoralis* from Washington to Colorado; Packard 1945; Macdougall-Shackleton et al. 2000), strong breeding site fidelity could maintain insularity and reduce gene flow across montane areas. Under this scenario, genetic similarities between taxa may be the result of recent divergence times and incomplete lineage sorting.

Broadly, we use the rosy-finch system to explore how patchy distributions affect genome-wide patterns of divergence. We use the above scenarios as a framework to investigate the evolutionary history of rosy-finches using high-resolution data from whole-genome sequences. Specifically, we ask if the isolation of alpine and high-latitude tundra habitat acts as a barrier to gene flow, and to what extent gene flow is connecting these populations.

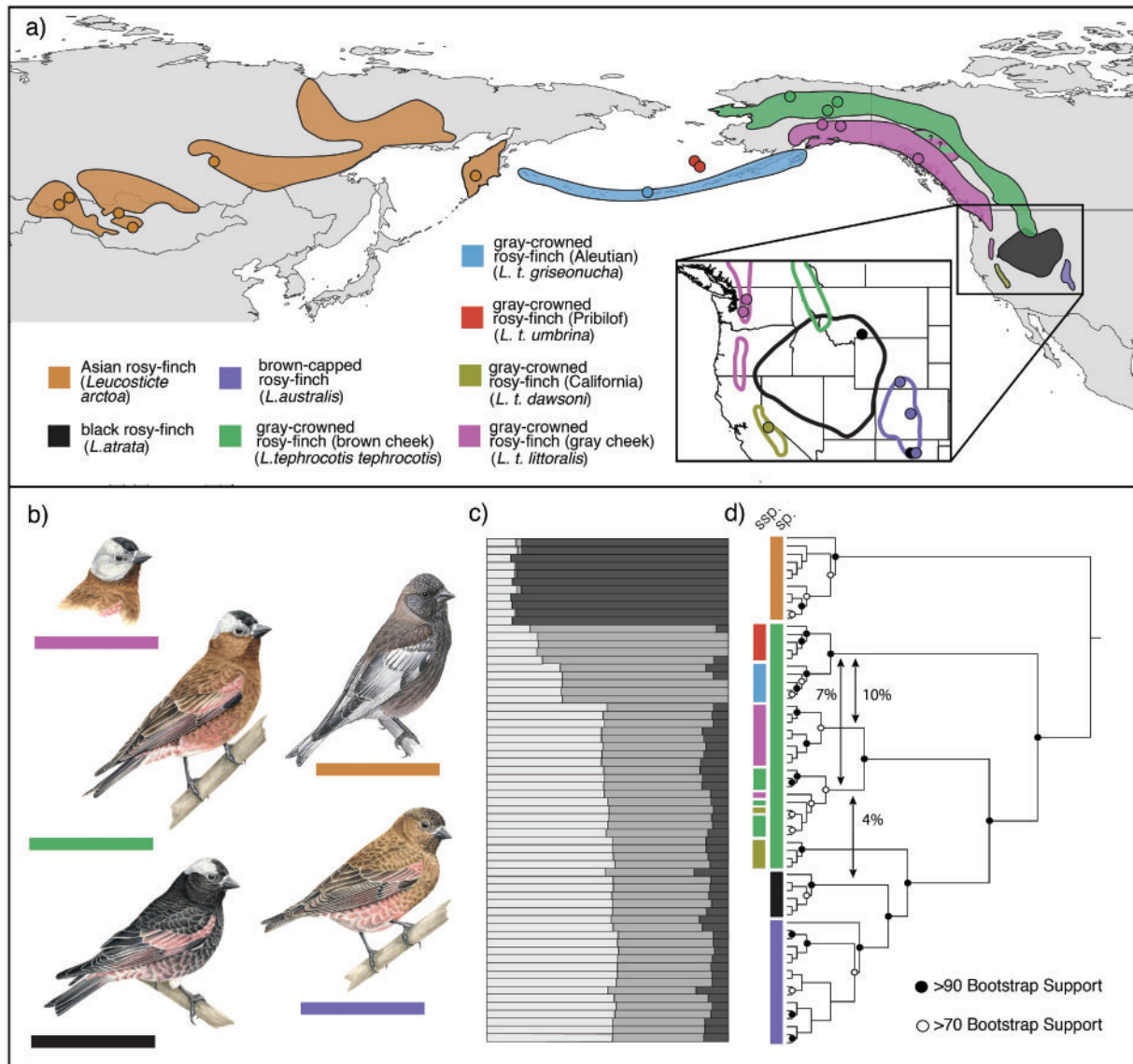


FIGURE 1. Overview of the rosy-finch species complex. a) Approximate breeding distribution of the rosy-finches across Asia and North America. Striped distribution indicates range overlap, and colored circles indicate sampling localities. Inset shows details of distribution across the lower 48 United States. b) Illustrations of five rosy-finch taxa. Color bars correspond to the legend in panel a). The Alaska island rosy-finches (not pictured) are similar to *L. t. littoralis* in that they possess a gray cheek, but exhibit larger, darker bodies. Rosy-finches from California (not pictured) are similar to *L. t. tephrocotis* but tawnier above. c) Population structure recovered from conStruct analysis. Bars represent individuals, and genomic admixture proportion represented by proportion of color within each bar. Plot shows result from a spatial model run with  $K = 3$ . All Asian rosy-finches are strongly clustered into one population, as well as all individuals from the Alaska islands. The remaining population includes all individuals from the North American mainland. d) SVDquartets phylogeny produced from 500,000 quartets with 200 bootstrap replicates including ~9,600,000 SNPs from the data matrix with no missing data filter and a 0.05 minor allele frequency. Open dots represent nodes with greater than 70% bootstrap support, and filled in dots represent nodes with greater than 90% bootstrap support. Clade bars indicate species and subspecies membership. Both black rosy-finches and brown-capped rosy-finches are monotypic and do not contain subspecies. Although we included four of the five subspecies of the Asian rosy-finch, we recovered no phylogenetic signal or population genetic structure, and do not differentiate them here. Our reconstruction suggests a paraphyletic gray-crowned rosy-finch with respect to the black and brown-capped rosy-finches, with strongly supported monophyly for Alaska island rosy-finches, as well as black and brown-capped rosy-finches. Arrows indicate the three highest values of  $F$  from ABBA-BABA analyses, indicating the percentage of the genome recovered as admixed. (Illustrations of mainland taxa: Liz Clayton Fuller; Asian rosy-finch: Lynx Edicions.)

We investigate differentiation across the species complex using a combination of principal component analyses, population assignment, and phylogenetic analyses, and use the results from these analyses to inform tests of introgression and models of evolutionary history.

By assessing multiple population genetic processes of potentially small or isolated populations, these patterns of divergence provide insights into the context under which speciation can occur among patchily distributed, but mobile, organisms.



## MATERIALS AND METHODS

*Sampling*

We sampled 68 individuals from across the range of the four rosy-finch species currently recognized by most authorities (Supplementary Table S1 available on dryad at <https://doi.org/10.5061/dryad.31zcrjdg9>; American Ornithologists' Union 1998; Clements et al. 2019), including four of the five Asian rosy-finch subspecies (*L. arctoa arctoa*,  $n=4$ ; *L. a. gigliolii*,  $n=2$ ; *L. a. sushkini*,  $n=4$ ; and *L. a. brunneonucha*,  $n=1$ ), five of the six gray-crowned rosy-finch subspecies (*L. tephrocotis tephrocotis*,  $n=8$ ; *L. t. littoralis*,  $n=9$ ; *L. t. griseonucha*,  $n=5$ ; *L. t. umbrina*,  $n=5$ ; *L. t. dawsoni*,  $n=5$ ), the black rosy-finch ( $n=6$ ), and the brown-capped rosy-finch ( $n=19$ ). Although this sampling is biased toward gray-crowned rosy-finches, it allows us to include as much of the described diversity of gray-crowned rosy-finches as possible, resulting in a more confident estimation of relationships and accurate detection of gene flow. We prepared genomic libraries using a modified Nextera Protocol (Illumina, Inc.). We diluted extracted DNA to a concentration of 2.5 ng/ $\mu$ L followed by tagmentation of 1  $\mu$ L gDNA. We indexed samples using the Nextera index kit (Illumina, Inc.) and amplified with KAPA HiFi Hot Start (Kapa Biosystems). We added primers with a reconditioning PCR and cleaned and size selected the products using an AMPure XP bead clean-up (Agencourt). We pooled libraries for sequencing on four lanes of an Illumina HiSeq X ten, with a library fragment size targeting 350 base pairs.

Details of the bioinformatic and filtering steps can be found on github ([https://github.com/erikrfunk/whole\\_genome\\_bioinformatics/blob/master/rosyfinch\\_notes.md](https://github.com/erikrfunk/whole_genome_bioinformatics/blob/master/rosyfinch_notes.md)). Briefly, we performed a quality assessment for each sample using fastQC v0.11.8 ([www.bioinformatics.babraham.ac.uk/projects/fastqc](http://www.bioinformatics.babraham.ac.uk/projects/fastqc)) before and after trimming reads. We used Trimmomatic v0.32 (Bolger et al. 2014) in paired-end mode to perform adapter removal and to trim reads using a sliding window when average Phred scores dropped below 20. We aligned all trimmed sequences to the House Finch (*Haemorrhous mexicanus*; Shultz 2017) genome using the BWA-MEM algorithm (Li and Durbin 2009) and sorted and indexed the alignment files using SAMtools v1.9 (Wysoker et al. 2009). We used the Genome Analysis ToolKit (GATK) v4.1.0.0 (McKenna et al. 2010) to call variants for each individual using HaplotypeCaller and merged all individual genomic variant call files into a cohort variant call file using CombineGVCFs (Poplin et al. 2017). We jointly genotyped individuals using the cohort variant call file with GenotypeGVCFs and applied the following hard filters to the resulting variants. We began by removing all insertions and deletions and set the following filter criteria according to the GATK recommended filter levels: QD < 2.0, FS > 60.0, MQ < 40.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. We also filtered out variants with a minor allele frequency less than 0.05, and that had an average depth less than one or greater than nine to remove variants possibly coming from paralogous loci. We constructed

multiple data matrices including one allowing for 25% missing data, which we call a 75% complete matrix, and one with no missing data filter. Due to the sensitivity of some programs to minor allele frequency, we also constructed an alternative dataset with the same filters as above, but with a minor allele frequency filter of 0.1 to examine the effects of rare alleles on analysis results (Linck and Battey 2019). We refer to the four datasets using a combination of the missing data proportion, and the minor allele frequency. These combinations include the 75% complete data matrix as "75p", and a matrix with no missing data filter as "0p." We also include our two minor allele frequency filters of 5% as "0.05maf," and 10% as "0.1maf." Finally, we assessed possible sibling relationships between sampled birds using the relatedness statistic from VCFtoolsv0.1.15 (none were found; Danecek et al. 2011). From these datasets, we calculated general population genetic statistics across 25,000 base-pair windows using VCFtools (Danecek et al. 2011), including nucleotide diversity ( $\pi$ ), Tajima's  $D$ , and Weir and Cockerham's pairwise  $F_{ST}$ , between all taxa.

*Principal Component Analysis*

To visualize total genomic variation, we plotted principal component analyses (PCAs) for three different groups of individuals. The first PCA contained all individuals whereas the two additional PCAs each contained a subset of individuals, including 1) only North American taxa, and 2) only mainland North American taxa, excluding taxa from the Aleutian and Pribilof islands. All analyses were conducted in R v3.6.3 using the package 'SNPRelate' version 1.12.2 (Zheng et al. 2012) and were repeated for each dataset.

*Population Assignment*

One major limitation of traditional approaches to assigning individuals to discrete populations is accounting for continuous variation. Patterns such as isolation by distance have the potential to create continuous variation in genetic data, where individuals geographically closer to one another may also appear more similar genetically (Wright 1943). Due to the discontinuous nature of sampling, population assignment in such cases might recognize many discrete populations, rather than one continuous population. Therefore, we used the R package 'conStruct' version 1.0.3 (Bradburd et al. 2018). This program can be run in both a spatial and a nonspatial mode. The nonspatial model is very similar to the program ADMIXTURE (Alexander et al. 2009; Bradburd et al. 2018); however, when given a distance matrix between all individuals, the program can run a spatial model to allow allele frequencies to vary by distance within a population, decreasing the chance of over-splitting populations. The support of a spatial model over a nonspatial model provides evidence of isolation by distance

(Bradburd et al. 2018). We used the R package ‘fields’ version 9.8.6 (Nychka et al. 2017) to construct a pairwise distance matrix for all individuals from locality coordinates. To more accurately estimate true avian dispersal distances from Asian to North American sampling locations, we adjusted the distances to track through a population in Alaska, before connecting to individuals from more southern populations.

We ran two independent chains of 10,000 iterations for both spatial and nonspatial models for two randomly sampled sets of 1,000,000 single nucleotide polymorphisms (SNPs). We evaluated trace plots of all posteriors to confirm that all runs converged. We modeled all number of populations from two to eight ( $K=2:8$ ), and used *conStruct*’s built-in function *x.validation* to evaluate the support for 1) the spatial versus nonspatial models and 2) best supported value of  $K$ . This function uses Monte Carlo cross-validation to perform model comparison. Briefly, for each model, 90% of the allele frequency data was randomly partitioned into a training set and used to estimate model parameters. A standardized mean log-likelihood was then calculated using the remaining 10% of the allele frequency table to generate a mean predictive accuracy (see Bradburd et al. 2018). In addition to predictive accuracy, we examined the contribution each added population provided, referred to in the R package as “layer contributions.”

### Phylogenetic Analysis

To evaluate the relationships between both individuals and potential populations, we constructed phylogenies of individual lineages using maximum likelihood in the program RAxML v.8.2.12 (Stamatakis 2014), and the quartet approach SVDquartets, implemented in PAUP\* v4.0a (Swofford 2003; Chifman and Kubatko 2014). The Asian rosy-finch was treated as the outgroup and was used to root all trees. Additional details of our phylogenetic methods are included in the [Supplementary material](#) (SM1.1 available on dryad) and can be found in the Dryad data repository (<https://doi.org/10.5061/dryad.31zcrjdg9>). Although other phylogenetic reconstruction approaches such as SNaQ allow for reticulations, they are limited in their scalability to larger datasets (Solís-Lemus and Ané 2016). Additionally, recent radiations of multiple lineages are likely to have upwards of 5–10 reticulations, reducing the feasibility of many network approaches. Although sub-setting data for these approaches are sometimes a viable option, it can reduce the resolution that a dataset provides. Species complexes consisting of multiple, recently diverged lineages often rely on the resolution of large datasets, such as whole genomes, in order to resolve relationships and population genetic structure. Both methods included here are capable of handling whole-genome sized datasets, and SVDquartets has been demonstrated to perform well even when systems have experienced gene flow (Long and Kubatko 2018).

### ABBA-BABA

Incomplete lineage sorting (ILS) and introgression from hybridization can produce gene trees that are incongruent with species trees. By investigating which of these processes may have occurred, or which is more predominant, we can better understand the broader evolutionary history of a group, including connectivity and gene flow versus recent isolation of populations. To evaluate the respective roles of ILS and introgression in the evolutionary history of rosy-finches, we calculated  $D$  and  $F$ -statistics using ABBA-BABA tests (Green et al. 2010; Martin et al. 2015).

ABBA-BABA analyses are constructed by examining the frequency of derived alleles in a rooted four-taxon phylogeny. These tests focus on genealogies that differ from the estimated underlying species topology. Under the hypothesis of ILS, each discordant topology is expected to occur at equal frequency, due to the stochastic nature of the lineage sorting process (ABBA = BABA). Alternatively, introgression from hybridization is likely to result in one discordant topology occurring at a much higher frequency due to the sharing of genetic material between two hybridizing taxa (ABBA < or > BABA). We constructed ABBA-BABA tests for all possible four-taxon topologies derived from well-supported clades from the 0p0.05maf SVDquartets analysis. We calculated the frequency of the derived allele within each taxon to incorporate differences among individuals. In the case where an entire taxon did not form a single clade, we included only individuals that together did form a clade and that represented a majority of individuals from that taxon. We then used a block-jackknifing approach to assess significant differences in the frequencies of each topology (Green et al. 2010). We used a  $Z$ -score cutoff value of 2.34 to assess significance at the 99% confidence level. We also used ABBA-BABA tests to estimate the proportion of the genome that is admixed between two populations as an  $F$ -statistic following Green et al. (2010), [Supplementary material 15](#) available on dryad (also see Martin et al. 2015). This approach compares the  $D$ -statistic calculated above with an expected value assuming complete admixture. The expected value is calculated by randomly assigning the individuals from the third taxon of the tree into two populations and substituting one as the second taxon. A  $D$ -statistic is then calculated for the randomized individuals and compared with the original  $D$ -Statistic. Subsequently, we used the results from these ABBA-BABA analyses to inform our construction of models of evolutionary history. All ABBA-BABA calculations were made using custom scripts written by Simon Martin ([https://github.com/simonhmartin/tutorials/tree/master/ABBA\\_BABA\\_whole\\_genome](https://github.com/simonhmartin/tutorials/tree/master/ABBA_BABA_whole_genome)).

### Modeling with Approximate Bayesian Computation

The alpine and tundra specialization of rosy-finches restricts their breeding distribution to small populations along high elevations or isolated northern islands.

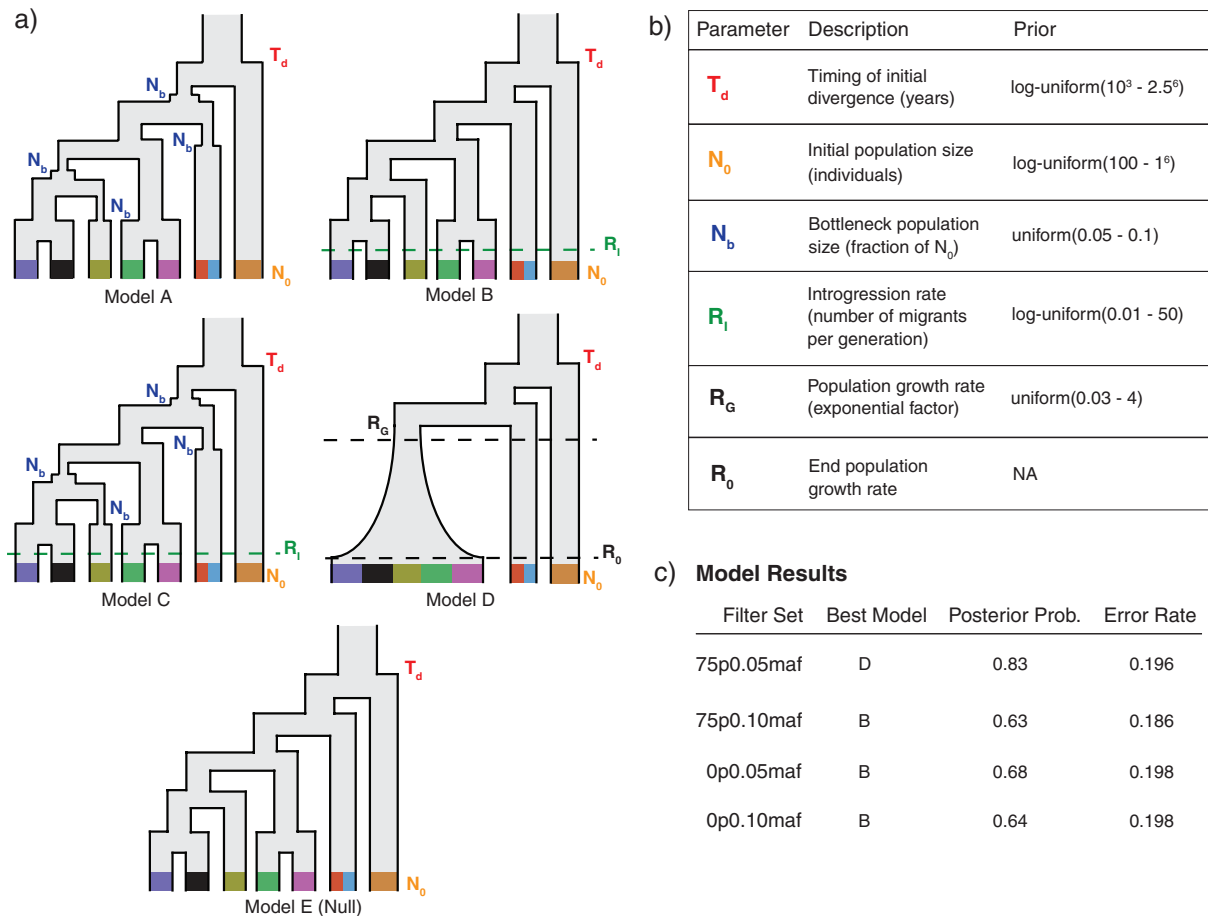


FIGURE 2. Models used in ABC analyses. a) Diagrams of the five models and their associated parameters. Models A and C include population bottlenecks at divergence events. Models B and C include symmetrical introgression between mainland taxa, as well as between *L. t. littoralis* and the Alaska island taxa. Model D considers all mainland taxa as a single lineage that underwent a population expansion. Model E acts as our null model and includes only divergence events. b) List of parameters and their description, including the values used for their prior ranges. c) Model results for all four filter sets.

As a result, their current distribution may be the result of repeated colonization of new mountaintops by relatively few individuals. This repeated colonization, as well as ILS and introgression (e.g. through annual recolonizing of breeding sites by mixed-origin flocks formed during the nonbreeding season), all have the potential to impact contemporary patterns of genomic variation. To test the roles that different processes played in the evolutionary history of rosy-finches, including population bottlenecks, population expansions, and introgression, we compared five different models of evolutionary history using approximate Bayesian computation (ABC) by simulating sequence data with msABC (Pavlidis et al. 2010; Smith et al. 2018). We constructed models that included one or more of these processes (Fig. 2a). The five models included (A) only population bottlenecks at divergence events, (B) only contemporary gene flow at the tips of the phylogeny, (C) both gene flow and population bottlenecks, and (D) a single North American mainland population that underwent population expansion. Model D was constructed to represent the previously proposed hypothesis of a single, widespread North American

lineage (Drovetski et al. 2009). We modeled bottlenecks on one of the two descending lineages to approximate a peripatric dispersal event. We also included a null model (E) with no events other than divergences. Support of this model would suggest ILS as a major contributor to observed genomic differences.

We evaluated each model using ABC by simulating multi-locus sequence data. We used a modified version of 'msABC' (Pavlidis et al. 2010; for version and modifications see Smith et al. 2018), which utilizes the coalescent-based software 'ms' (Hudson 2002). We selected up to five of the most deeply sequenced individuals from each population, informed from phylogenetic analysis. In the case an entire taxon did not form a single clade, we selected only individuals that together did form a clade and that represented the majority of individuals from that taxon. We filtered the total dataset to construct a set of short loci appropriate for these analyses, targeting regions of the genome that were between 250 and 500 base pairs long with at least  $4\times$  coverage, and were present in more than three of the five individuals selected per population. To reduce the likelihood of using linked loci, we required



a minimum distance of 10,000 base pairs between each locus (Balakrishnan and Edwards 2009). We included two sequences per individual so that both alleles were included for heterozygotes. These filters resulted in 7423 loci.

For each model, we fixed the order of divergence events to match that produced by the SVDquartets phylogeny. We set a broad uniform prior on the root of the phylogeny from 10,000 ya to 2.5 Ma and set the timing of each subsequent event as conditional on the previous event. We estimated the mutation rate under a uniform prior from  $1 \times 10^{-9}$  to  $3 \times 10^{-9}$  per site per year, so that the average value was equal to the estimated mutation rate reported for *Geospiza magnirostris* (Rands et al. 2013). All population sizes were estimated under a log-uniform prior from 100 to  $1 \times 10^6$  individuals. The list of events and their priors is outlined for each model in Figure 2b. We simulated 50,000 datasets for each model and used msABC to calculate summary statistics for both the simulated and empirical datasets. Because we included fewer samples from the Asian rosy-finch, we prioritized modeling the evolutionary history of the North American taxa and did not include summary statistics for the Asian rosy-finch. Summary statistics included the average and variance of nucleotide diversity ( $\pi$ ), Watterson's theta, number of segregating sites, fixed differences, private alleles, Tajima's  $D$ , and pairwise  $F_{ST}$  (Hudson et al. 1992). We performed model selection with a machine-learning approach implementing random forests using the R package 'abcrf' version 1.8 (Pudlo et al. 2015). This method uses the simulated data to train decision trees that provide a model prediction. When applied to summary statistics from an observed dataset, each tree then provides a model vote, where the model with the most votes is deemed the best suited model. The posterior probability of the selected model is then calculated in a separate, second step, using an additional random forest.

## RESULTS

### Sampling

Sequencing produced a total of  $3.6 \times 10^9$  raw reads across all 68 individuals, resulting in an expected depth of raw coverage of approximately

$6.5 \times$  per individual. One individual (*L. australis*) was dropped from further analyses due to demultiplexing ambiguities (Supplementary Table S1 available on dryad). Approximately 98% of reads mapped to the House Finch genome for each individual, indicating reasonable alignment. Two individuals were dropped due to low total coverage (Supplementary Table S1 available on dryad), resulting in a range of coverage from  $1.25 \times$  to  $6.62 \times$  for all included variants. In our sampling, a single individual (RF8) appeared to be morphologically intermediate between a black rosy-finch and a brown-capped rosy-finch. Although an in-depth description of this individual as a potential hybrid is warranted, it is beyond the scope of this paper. In order to limit bias toward detecting introgression, we dropped this individual from ABBA-BABA, ABC, and phylogenetic analyses. The remaining individuals were used to generate four datasets that varied in both missing data and minor allele frequency. The smallest dataset was 96,803 SNPs (75p0.1maf), whereas the largest dataset was 9,819,646 SNPs (0p0.05maf).

Sampled populations exhibit only minor differences in nucleotide diversity (Table 1), with the lowest diversity occurring in the population of individuals from the Alaska islands ( $\pi = 0.0006$ ). Estimates of Tajima's  $D$  are sensitive to the minor allele frequency filter used, which prohibits their meaningful interpretation, but the Asian rosy-finch is consistently recovered as possessing the highest Tajima's  $D$  ( $D = 1.5$ ; Table 1). Pairwise  $F_{ST}$  varies greatly across comparisons, ranging from 0.01 to 0.39 (Supplementary Table S2 available on dryad), with the strongest differentiation occurring between the Asian rosy-finch and all other populations, followed by the Alaska island populations and all other populations (Supplementary Table S2 available on dryad).

### Principal Component Analysis

All datasets produced similar PCAs, and we present here the results from the 75p0.05maf dataset. The PCA including all 65 of the retained individuals strongly separates all North American individuals from Asian individuals along PC1 (20.1% of variation; Fig. 3a). North American individuals are continuously distributed across PC2 (5.5% of variation) with little differentiation between currently described taxa. The PCA including

TABLE 1.  $\pi$  and Tajima's  $D$  calculated for all currently recognized taxa

Taxon	$\pi$ —windowed <sup>a</sup>	$\pi$ —sites	TajD—windowed <sup>a</sup>
<i>Leucosticte arctoa</i>	0.00078028	0.211322	1.58631
<i>Leucosticte tephrocotis umbrina</i> and <i>griseonucha</i>	0.00055124	0.139382	0.925215
<i>Leucosticte tephrocotis tephrocotis</i>	0.00065504	0.175463	0.464404
<i>Leucosticte tephrocotis littoralis</i>	0.00071449	0.177705	0.404722
<i>Leucosticte tephrocotis dawsoni</i>	0.00071543	0.158612	0.611621
<i>Leucosticte atrata</i>	0.00083945	0.180744	0.36156
<i>Leucosticte australis</i>	0.00081807	0.183758	0.640382

Note: <sup>a</sup> Average across a 25-kb window.

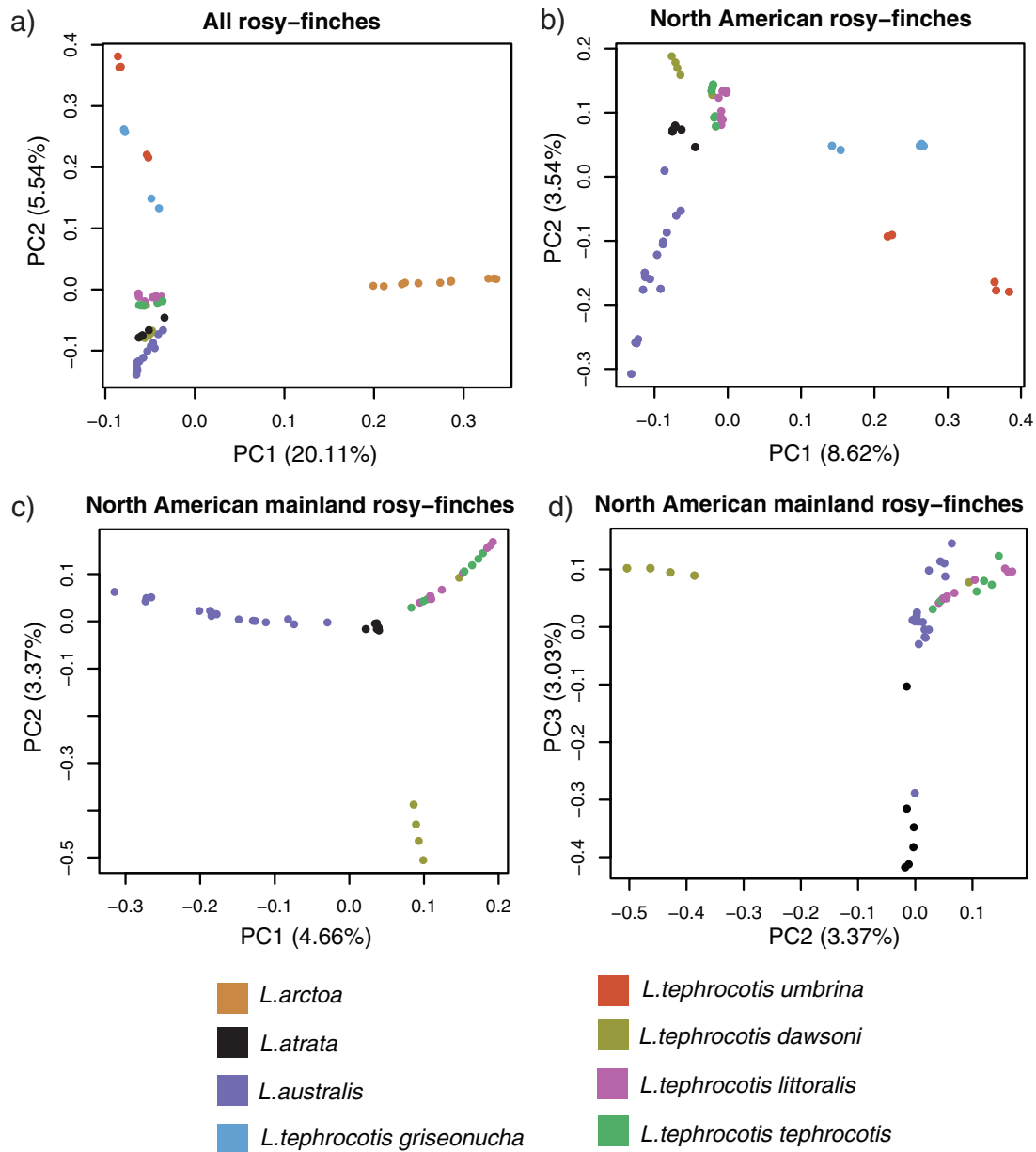


FIGURE 3. PCAs summarizing total genomic variation of 6,137,268 SNPs from the 75% complete data matrix and 0.05 minor allele frequency filter. a) All SNPs for all individuals, showing strong separation between Asia and North American individuals. b) North American individuals only, showing strong separation between Alaska island populations and North American mainland populations. c) North American mainland individuals only, showing continuous variation across PC1. Four unrelated gray-crowned rosy-finch individuals from California are separated along PC2. d) North American mainland individuals, showing PC2 versus PC3.

only North American individuals strongly separates individuals from both the Aleutian and Pribilof (Alaska) islands from North American mainland individuals along PC1 (8.62% of variation; Fig. 3b). The second PC axis (3.54% of variation) weakly differentiates North American mainland taxa, however, variation still appears continuous. The PCA containing only North American mainland individuals shows continuous variation across PC1 (4.66% of variation) including all currently described North American species (Fig. 3c). Although these secondary axes appear to represent

only small amounts of total variation, small values are typical for genome-scale data at shallow evolutionary timescales (Toews et al. 2016a; Beckman et al. 2018). Four individuals from the California subspecies of gray-crowned rosy-finch (*L. t. dawsoni*) are strongly separated along the second PC axis (3.37% of variation; Fig. 3c). These individuals do not appear to be related, a factor that can result in this pattern. All individuals of the black rosy-finch are clustered centrally along the first and second PC axes, but show continuous variation along the third axis (3.03% of variation, Fig. 3d).



### Population Assignment

All datasets produced consistent results, and we present here the result for the 75p0.1maf dataset. For all values of  $K$ , conStruct plots show little variation between spatial and nonspatial models; however, model comparison using cross-validation recovers support for the use of a spatial model. We used both gains in predictive accuracy and contributions of additional clusters to select the most appropriate value of  $K$ . We find support for a  $K$  of 3 due to the smaller improvement of predictive accuracy from cross-validation (Supplementary Fig. S1a available on dryad) and diminishing contributions of higher values of  $K$  (Supplementary Fig. S1b available on dryad). Results from the three-cluster spatial model (Fig. 1c) identify a well-defined Asian rosy-finch cluster, along with a well-defined cluster that includes the Aleutian and Pribilof islands. The third cluster includes all remaining individuals, corresponding to North American mainland and suggests very little variation among individuals within the North American mainland populations. Higher values of  $K$  (Supplementary Fig. S1c available on dryad) differentiated the two different Alaska island populations but continued to indicate a single North American mainland genetic cluster. When Asian rosy-finch samples are removed from the analysis, the remaining genetic clusters are congruent with the results presented in Figure 1c, recovering only island, and mainland clusters. This is also true when only samples from mainland North America are included in the analysis; construct does not detect population structure among mainland North America samples. Although including samples collected during the nonbreeding season has the potential to impact the spatial model, these individuals were recovered with a similar admixture proportion as conspecific individuals in both the spatial and nonspatial models. These nine samples (Supplementary Table S1 available on dryad) were identified on the basis of phenotype and were generally only minimal distances from their breeding ranges (the furthest being black rosy-finches from New Mexico, which most likely breed in Utah or Wyoming).

### Phylogenetic Analysis

All phylogenetic analyses recover full support for a monophyletic Asian rosy-finch (Fig. 1d). The relationships recovered for North American rosy-finches varied among datasets, and among tree construction methods (Supplementary Figs. S2a and S3 available on dryad). Although variation is present, many strongly supported consistencies exist that allow for generalization of the relationships of major clades. The RAxML tree from the 0p0.05maf dataset produced the largest number of supported nodes and did not conflict with any of the well-supported nodes of major clades recovered in other RAxML analyses. Although there was some variation in the SVDquartets topologies (discussed

below) the SVDquartets tree from the 0p0.05maf dataset was consistent with the RAxML results, and added further resolution with additionally supported nodes, fully resolving the backbone of the tree. The placement of a single black rosy-finch (RF37, MSB:30447) was variable across analyses; however, this appears likely to be the result of low overall coverage and its placement is consistent in the two 0p trees discussed above. We view the broad congruence between these 0p trees as support for a single best phylogenetic hypothesis (Fig. 1d) and discuss that topology here.

Within North American rosy-finches, the island populations of gray-crowned rosy-finches were recovered as sister to all mainland rosy-finches. Additionally, within the clade of island individuals, the two currently recognized subspecies from the Aleutian and Pribilof islands (*L. t. griseonucha* and *L. t. umbrina*, respectively) are recovered as reciprocally monophyletic. As a species, the gray-crowned rosy-finch is recovered as paraphyletic, with all black and brown-capped individuals nested within it. Well-supported clades are also recovered that correspond to some of the currently described mainland taxa, but not all (Fig. 1d). This includes fully supported clades corresponding to the black rosy-finch and to the brown-capped rosy-finch. The two widespread North American mainland subspecies of the gray-crowned rosy-finch (*L. t. tephrocotis* and *L. t. littoralis*) are all recovered as closely related; however, the clade also included one individual from the California subspecies (*L. t. dawsoni*; Fig. 1d).

Although one California individual is recovered within the rest of the gray-crowned individuals, the placement of the four remaining California individuals is less certain. SVDquartets phylogenies constructed from the different minor allele frequency datasets produced well supported, but conflicting topologies. Relationships reconstructed with the 0.05maf dataset (Supplementary Fig. S2a available on dryad) recovered black and brown-capped rosy-finches as sister, with the California individuals as sister to them, whereas the 0.1maf dataset (Supplementary Fig. S2b available on dryad) recovered the California individuals as sister to black rosy-finches.

### ABBA-BABA

ABBA-BABA analyses reveal significant signals of introgression between multiple pairs of taxa (Supplementary Table S3a available on dryad). Introgression was detected between the Alaska island individuals, and the four mainland taxa, including the black rosy-finch, and the two widely distributed subspecies of gray-crowned rosy-finch (*L. t. tephrocotis* and *L. t. littoralis*). A small, but significant level of introgression was also recovered between the brown-capped rosy-finch and the Alaska island populations. Introgression appears to be widespread across mainland taxa, with a significant result in every

comparison. Values of  $D$  from these comparisons were higher between geographically proximate taxa. For example, values of  $D$  were higher between the black rosy-finch and the gray-crowned rosy-finch than between the brown-capped rosy-finch and the gray-crowned rosy-finch (Supplementary Table S3 available on dryad). The California individuals demonstrate higher levels of introgression with the black rosy-finch than with any other group, despite the California population's current classification as a subspecies of the gray-crowned rosy-finch. Both the widespread North American mainland subspecies (*L. t. tephrocotis* and *L. t. littoralis*) inhabit allopatric mountain ranges separated by broad interior river valleys in Alaska, and southern British Columbia. To what extent, these two subspecies may come into contact at the headwaters of these major rivers in the southern Yukon Territory and northern British Columbia is currently unknown, but the two taxa demonstrate different patterns of introgression with respect to more outlying taxa. We find that the gray-cheeked subspecies (*L. t. littoralis*) exhibits slightly higher rates of introgression with the Alaska island taxa, whereas the brown-cheeked subspecies (*L. t. tephrocotis*) exhibits higher rates with the other mainland taxa (brown-capped and black rosy-finches: *L. atrata* and *L. australis*). These results suggest higher rates of introgression between taxa that share similar plumage morphology. More specifically, the island taxa and *L. t. littoralis* both possess gray cheek feathers, whereas *L. t. tephrocotis* and the other mainland taxa lack gray cheek feathers, and instead possess brown or black cheeks.

Results from the  $F$ -statistic calculations of ABBA-BABA analyses suggest the highest admixture proportion is between the Alaska island populations, and the gray-cheeked subspecies of the widespread North American subspecies of gray-crowned rosy-finch (*L. t. littoralis*), with approximately 10% of their genome recovered as admixed (Fig. 1d). Admixture proportion is slightly lower between the island population and the brown-cheeked subspecies of the widespread North American subspecies of gray-crowned rosy-finch (*L. t. tephrocotis*), at 7%, and approximately 3% between the black rosy-finch and the two widely distributed gray-crowned rosy-finch taxa (*L. t. tephrocotis*, and *L. t. littoralis*). Results from all comparisons can be found in Supplementary Table S3b available on dryad.

#### Modeling with Approximate Bayesian Computation

We evaluated five models of evolutionary history using ABC to examine the role of gene flow versus that of population size changes (bottlenecks or expansion; Fig. 2). Model comparison using random forests recovered the highest support for model B (introgression only) for all filter sets except 75p0.05maf, which favored model D (rapid population expansion; Fig. 2c). We report results here for both 0.05maf filter sets. The 0p0.05maf filter set favored model B with a posterior probability of 0.68 and an out-of-bag prior error rate of 0.198.

Of the 1000 model votes generated by the random forest, model B received 418 votes, followed closely by model C (introgression and population bottlenecks) with 358 votes. Model D received 116 votes whereas model E received 76 votes and model A received 32 votes. The 75p0.05maf filter set favored model D with a posterior probability of 0.83 and an out-of-bag error rate of 0.196. Of the 1000 model votes generated by the random forest, model D received 356 votes, followed by model B (introgression only) with 312 votes. Model C received 240, model E received 60, and model A received 32 votes.

#### DISCUSSION

We employed a population genomic and phylogenomic approach to examine patterns of differentiation and gene flow among all major rosy-finch taxa, which are high elevation sky island and arctic tundra specialists. We recovered strongly supported clades corresponding to the majority of currently described taxa, but found evidence from multiple analyses of widespread introgression between many pairs of taxa despite the highly fragmented nature of rosy-finch breeding habitat. Our results indicate that patchy distributions do not act as barriers to gene flow in species with a high capacity for dispersal, like the rosy-finches. Additionally, our findings suggest that phylogenetic signal and, more generally, divergence, can occur across patchily distributed organisms, despite high dispersal and introgression. Distinguishing evolutionary processes can be difficult, especially in nonmodel systems with complex natural and evolutionary histories. However, understanding the effects of these processes and their interactions can provide insight into the broader context in which divergence and speciation occur. Our results suggest that at least some portion of the rosy-finch genome has been more strongly affected by drift or selection when compared with background levels of genome differentiation that have likely been affected by gene flow.

#### Differentiation and Introgression across Rosy-Finches

All analyses that examine divergence strongly support both the Asian rosy-finch, and rosy-finches from the Alaska islands, as distinct lineages (Figs. 1c,d and 3). These two taxa also exhibit the highest pairwise  $F_{ST}$  values in all comparisons (Supplementary Table S2 available on dryad), further supporting the strength of their differentiation. Divergence of the Alaska island populations could be occurring for multiple reasons, including localized selection, and/or increased rates of drift caused by small population sizes relative to the mainland populations. Additionally, this signal of divergence may have developed ancestrally in a Beringian glacial refugium, as has been suggested for the rock ptarmigan (*Lagopus muta*), rock sandpiper (*Calidris*

*ptilocnemis*), and Common Raven (*Corvus corax*; Winker et al. 2002; Pruett and Winker 2008). Despite including four of the five subspecies of the Asian rosy-finch, we find no support for differentiation among these taxa, mirroring results from both PCA and conStruct for mainland North America rosy-finches. This result suggests further investigation with more sampling among the Asian populations is warranted.

Although results were consistent with respect to the divergence of Asian rosy-finches and Alaska island rosy-finches, patterns of differentiation in mainland North America vary by analysis. Both PCA and conStruct recover all mainland North America individuals as a single, continuously varying group. When running conStruct's spatial model, which allows allele frequencies to vary across space, these individuals contain roughly the same admixture proportions. This same pattern emerges in the PCA of mainland individuals as a single, nearly continuous line across PC1 (with the exception of some California individuals, discussed below). The order of these individuals also corresponds broadly with geography, suggesting this pattern may be driven in part by isolation by distance. This conclusion is supported by the spatial model in conStruct (Fig. 1c), as well as in an increase in the *D*-statistic between more geographically proximate populations (Supplementary Table S3 available on dryad).

The results from population clustering and PCA are reflective of a pattern in which the genomic backgrounds of mainland individuals have largely been homogenized among neighboring populations. However, consistently recovered phylogenetic clades within mainland North America (Fig. 1d) support two additional conclusions: 1) Some, likely few, regions of the genome contain signals of differentiation among mainland North American taxa and 2) the signal contained in these divergent loci is congruent with broader patterns of morphological variation. Given the short time since divergence, the congruence between phylogeny and morphology suggests that the genomic regions underlying morphological differences might be responsible for driving phylogenetic relationships. Additionally, given the broadly homogenized genomic background, these loci are likely few in number. Although genomic similarities across taxa may be the result of ILS, ABBA-BABA analyses support the presence of introgression in nearly all comparisons. Furthermore, although other demographic factors or ancestral population structure can affect ABBA-BABA tests (Eriksson and Manica 2012), ABC model results highlight introgression as a probable primary factor in producing the documented genomic similarities across taxa. That said, ABC model results do not completely rule out population expansion. Like other patchily distributed organisms with high dispersal capabilities (Zarza et al. 2016), our data provide support for the presence of gene flow across mountaintop sky islands in rosy-finches. Taken together, our results also support a growing body of literature highlighting the prevalence

of gene flow among recently diverged, or currently diverging lineages (Pinho and Hey 2010; Ellegren et al. 2012; Campbell et al. 2018).

The maintenance of phylogenetic signal in rosy-finches (i.e. Fig. 1c,d) poses interesting questions more generally about the demographic parameters that may allow populations to diverge, despite substantial levels of gene flow. One possible explanation could be the presence of strong metapopulation dynamics, including numerous, small populations. With small effective population sizes, novel alleles are more likely to increase in frequency due to genetic drift and/or localized selection. An intermediate level of gene flow across these populations would result in a detectable signal of introgression, but not one that is so strong that it masks the signal of divergence. This process of drift outweighing gene flow across patchy population distributions has been demonstrated in species with both short dispersal distances (González-Serna et al. 2018; Cameron et al. 2019) and longer dispersal distances (Llorens et al. 2017). In rosy-finches, previous studies have also revealed strongly skewed sex ratios (French 1959a), with males outnumbering females by as much as 6:1. Small effective population sizes caused by skewed sex ratios would enhance signals of divergence by allowing private or novel alleles to increase in frequency through drift. These hypotheses could be examined more directly in a population genomic framework with the inclusion of more densely sampled populations and estimation of demographic parameters. Although beyond the scope of this study, determining the specific genomic regions that provide phylogenetic resolution despite extensive gene flow may allow for a better understanding of the degree to which certain traits can drive divergence and speciation. Given the detection of both divergence and gene flow in this study, rosy-finches may be a system that is well suited to explore the role of localized selection, and further contrast current models of speciation-with-gene-flow (Nosil 2008; Pinho and Hey 2010), which highlights the role of selection in divergence, from speciation-with-bouts-of-gene-flow (Zarza et al. 2016), in which the role of selection is less clear.

#### *Considerations from high-resolution sequence data*

The most recent investigation of rosy-finch genetics reported widespread similarity among all North American forms, with the only supported divergence occurring between the Asian and North American taxa (Drovetski et al. 2009). However, this study was limited in the number of loci. With the application of whole-genome sequence data, we provide additional resolution to evaluate this hypothesis.

Similar to Drovetski et al. (2009), we find support for a strong divergence between Asian rosy-finches and North American rosy-finches. We find additional evidence of divergence between populations of rosy-finch on the Aleutian and Pribilof islands, and the rest of the North



American mainland. Broadly, we also find support for substantial levels of connectivity among the mainland populations and a pattern of isolation by distance. Despite evidence of gene flow, however, phylogenetic analyses support additional clades within the North American individuals (Fig. 1d). Many of these clades correspond to morphological or ecological differences previously used in describing species boundaries. It is possible that the support for these finer-scale clades is being driven by loci responsible for morphological or ecological differences, which are only captured when employing high-resolution whole-genome sequencing. Although species designation is likely to vary depending on the definition applied, a phylogenetic species concept recognizes at least four species from our data, with fully supported monophyly of the Asian rosy-finch, the black rosy-finch, the brown-capped rosy-finch, and the Alaska island rosy-finches (Supplementary SM1.2 available on dryad). This taxonomy omits a reduced and paraphyletic gray-crowned rosy-finch, but its paraphyly is unsurprising given the recent time scale during which these taxa likely diverged, with paraphyly often preceding monophyly during divergence (Neigel and Avise 1986). Given the gene flow we have found, it is uncertain whether gray-crowned rosy-finches might eventually become monophyletic across most of the genome. Currently accepted taxonomy in North America, which is based on the biological species concept and used to structure our sampling, seems to be supported from our results, albeit with a paraphyletic gray-crowned rosy-finch. This, however, is a classic pattern among biological species, termed “budding” by Haffer (1992, compare Fig. 3b with our Fig. 1d).

In general, phylogenetic support for the above clades is consistent across datasets when considering well-supported nodes. However, the placement of one of the clades (California gray-crowned rosy-finches; Supplementary Fig. S2 available on dryad) was less certain. The addition of rare alleles allowed by the lower minor allele frequency filter results in conflict with respect to four of the five individuals from California. These same individuals are also recovered as distinct in the PCA of mainland rosy-finches. Although nearly all variation is continuous along PC1, the four *L. t. dawsoni* individuals are separated along PC2 (Fig. 3c), lending some support to the conclusion these individuals form a distinct lineage. The California population appears to be one of the more geographically isolated populations of the gray-crowned rosy-finch, with potential connections to the gray-crowned populations to the north and the black rosy-finch to the east (French 1959b; Johnson 1975, 2002). Given the higher *D*-statistic between the black rosy-finch and California rosy-finches compared with the remaining gray-crowned populations, introgression may be responsible for generating the conflict in topology (Maddison 1997). Although additional sampling from across the range of California is likely necessary to resolve the uncertainty, differential rates and timing of introgression between populations, selection pressures,

or un-modeled demographic processes may be driving the differences between the two datasets.

Our results highlight an important consideration when interpreting whole-genome sequence data. Higher-resolution genomic data make possible the detection of even minor differences between populations. However, the evolutionary significance of these differences should be thoroughly explored. Furthermore, how these types of minor differences may affect analyses, or how results might vary based on the statistical approach of the tool (e.g. clustering approaches like PCA vs. generative approaches like phylogenetic models) remains unclear and is worth further exploration using simulations.

Although many phenotypic differences may play a role in speciation by facilitating divergence and reproductive isolation, divergence may result from processes other than directional selection (e.g. drift). It then becomes important to consider specifically how genomic regions differ, and the broader functions of these genomic regions. In the context of rosy-finches, distinct populations, like those on the Alaska islands, demonstrate a number of potentially adaptive differences (in addition to small population sizes relative to mainland populations) that might be driving divergence. In addition to size and color differences between mainland individuals, the island populations inhabit lower elevations with less seasonal movement (Macdougall-Shackleton et al. 2000), all of which could reduce gene flow with other populations and result in increased drift and local adaptation. Although mainland populations are known to differ in morphological characters such as coloration, it is uncertain if there are additional, potentially adaptive differences between these populations. These questions could be examined by searching for outlier loci that are under selection, but this is beyond the scope of the current study.

More generally, at shallow time scales or in hybridizing lineages, signals of divergence may be limited to few loci (or many loci with small effect) and, as a result, it is important to consider both how the information content of these loci affects analyses, and what the biological importance of these loci are in the context of population divergence and speciation (Toews et al. 2016b; Campagna et al. 2017; Marques et al. 2017; Ravinet et al. 2017). Furthermore, if relatively few loci exhibit signals of divergence at shallow time scales, they may be particularly sensitive to combinations of missing data, and filters, such as minor allele frequency. This sensitivity may explain some of the differences in results recovered here among datasets with different filtering parameters. Filter settings will have profound and direct effects on many analyses, including standard population genetic summary statistics such as nucleotide diversity and Tajima's *D* (Linck and Battey 2019). For example, in rosy-finches, altering filter settings can result in a completely flipped sign of Tajima's *D*, precluding any meaningful interpretation of selective or demographic processes shaping this system. This problem may be especially pronounced when analyzing lower coverage



sequence data as was used here. As a result, we do not make any strong conclusions based on these statistics.

### CONCLUSIONS

We find support for widespread gene flow among populations of a songbird whose breeding range is restricted to alpine and arctic tundra, despite recovering well-supported clades that largely reflect contemporary taxonomy (Fig. 1d). We pair ABBA-BABA analyses with ABC models to disentangle multiple evolutionary processes that have the potential to produce similar genomic patterns and find further support for our conclusion of widespread introgression. Overall, these results suggest a signal of divergence in some regions of the genome that outweighs the homogenizing signal of gene flow, likely resulting from a balance of selection, drift, and gene flow among numerous, small populations. Although we do not attempt to separate the effects of selection from drift in this paper, our results are similar to other studies that have identified patterns of population divergence attributed to drift resulting from patchy population distributions (Llorens et al. 2017; Sánchez-Montes et al. 2018; Cameron et al. 2019). Although gene flow is likely common in the early stages of speciation, the application of high-resolution genome sequencing makes it possible to detect signals of divergence for species with patchy distributions, and complex natural histories.

### DATA AVAILABILITY

Raw sequence data are available for download from the NCBI Sequence Read Archive (SRA) (BioProject PRJNA659436). Sequence alignments and phylogenetic trees are available on TreeBASE: <http://purl.org/phylo/treebase/phyloids/study/TB2:S26830>.

### SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.31zcrjdg9>

### ACKNOWLEDGMENTS

We thank Stacey Smith, Ethan Linck, members of the Taylor lab, and three anonymous reviewers for comments on previous versions of this manuscript. We also thank the University of Washington Burke Museum (UWBM), the Museum of Southwestern Biology (MSB), the Museum of Vertebrate Zoology (MVZ), the Yale Peabody Museum (YPM), the University of Alaska Museum (UAM), The KU Biodiversity Institute and Natural History Museum (KU), and the Denver Museum of Nature and Science (DMNS) for providing tissue and

blood samples. We thank Chris Smith for his assistance with ABC analyses.

### FUNDING

This work was supported by National Geographic WW-202R-17 (K.C.R.), the Beverley Sears Research Grant and the Alex Singer Memorial Grant from the University of Colorado Boulder, and a Graduate Student Research Award from the Society of Systematic Biologists (E.R.F.).

### REFERENCES

- Alexander D.H., Novembre J., Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655–1664.
- American Ornithologists' Union. 1931. Check-list of North American birds. Lancaster (PA): AOU.
- American Ornithologists' Union. 1983. Check-list of North American birds. Lawrence (KS): AOU.
- American Ornithologists' Union. 1998. Check-list of North American birds. Washington (DC): AOU.
- Balakrishnan C.N., Edwards S.V. 2009. Nucleotide variation, linkage disequilibrium and founder-facilitated speciation in wild populations of the zebra finch (*Taeniopygia guttata*). *Genetics.* 181:645–660.
- Beckman E.J., Benham P.M., Cheviron Z.A., Witt C.C. 2018. Detecting introgression despite phylogenetic uncertainty: the case of the South American siskins. *Mol. Ecol.* 27:4350–4367.
- Bolger A., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–2120.
- Bradbud G.S., Coop G.M., Ralph P.L. 2018. Inferring continuous and discrete population genetic structure across space. *Genetics.* 210:33–52.
- Cameron A.C., Page R.B., Watling J.I., Hickerson C.A.M., Anthony C.D. 2019. Using a comparative approach to investigate the relationship between landscape and genetic connectivity among woodland salamander populations. *Conserv. Genet.* 20:1265–1280.
- Campagna L., Repenning M., Silveira L.F., Suertegaray Fontana C., Tubaro P.L., Lovette I.J. 2017. Repeated divergent selection on pigmentation genes in a rapid finch radiation driven by sexual selection. *Sci. Adv.* 3:e1602404.
- Campbell C.R., Poelstra J.W., Yoder A.D. 2018. What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biol. J. Linn. Soc.* 124:561–583.
- Chifman J., Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics.* 30:3317–3324.
- Claramunt S., Derryberry E.P., Remsen J.V., Brumfield R.T. 2012. High dispersal ability inhibits speciation in a continental radiation of passerine birds. *Proc. R. Soc. B.* 279:1567–1574.
- Clements J.F., Schulenberg T.S., Iliff M.J., Roberson D., Fredericks T.A., Sullivan B.L., Wood C.L. 2019. The eBird/Clements checklist of birds of the world: v2019. Available from: <http://www.birds.cornell.edu/clementschecklist/download>.
- Cruikshank T.E., Hahn M.W. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Mol. Ecol.* 23:3133–3157.
- Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E., Lunter G., Marth G.T., Sherry S.T., McVean G., Durbin R., Group 1000 Genomes Project Analysis. 2011. The variant call format and VCFtools. *Bioinformatics.* 27:2156–2158.
- Dickinson E.C., Christidis L., editors. 2014. The Howard and Moore complete checklist of the birds of the world, 4th ed. Vol 2: Passerines. Eastbourne, England: Aves Press.
- Drovetski S.V., Zink R.M., Mode N.A. 2009. Patchy distributions belie morphological and genetic homogeneity in rosy-finches. *Mol. Phylogenet. Evol.* 50:437–445.

- Eimes J.A., Bollmer J.L., Whittingham L.A., Johnson J.A., Van Oosterhout C., Dunn P.O. 2011. Rapid loss of MHC class II variation in a bottlenecked population is explained by drift and loss of copy number variation. *J. Evol. Biol.* 24:1847–1856.
- Ellegren H., Smeds L., Burri R., Olason P.I., Backström N., Kawakami T., Künstner A., Mäkinen H., Nadachowska-Brzyska K., Qvarnström A., Uebbing S., Wolf J.B.W. 2012. The genomic landscape of species divergence in *Ficedula flycatchers*. *Nature*. 491:756–760.
- Eriksson A., Manica A. 2012. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. *Proc. Natl. Acad. Sci. USA*. 109:13956–13960.
- French N.R. 1959a. Life history of the Black Rosy Finch. *Auk*. 76:159.
- French N.R. 1959b. Distribution and migration of the Black Rosy Finch. *Condor*. 61:18–29.
- Gaither M.R., Bernal M.A., Coleman R.R., Bowen B.W., Jones S.A., Simison W.B., Rocha L.A. 2015. Genomic signatures of geographic isolation and natural selection in coral reef fishes. *Mol. Ecol.* 24:1543–1557.
- González-Serna M.J., Cordero P.J., Ortego J. 2018. Using high-throughput sequencing to investigate the factors structuring genomic variation of a Mediterranean grasshopper of great conservation concern. *Sci. Rep.* 8:1–14.
- Green R.E., Krause J., Briggs A. W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai W., Fritz M.H., Hansen N.F., Durand E. Y., Malaspina A., Jensen J. D., Marques-Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A. 2010. A draft sequence of the neandertal genome. *Science*. 328:710–722.
- Haffer J. 1992. The history of species concepts and species limits in ornithology. *Bull. Bull. B. O. C. Centenary Suppl.* 112A:107–158.
- Howell T.R., Paynter R.A. Jr., Rand A.L. 1968. Subfamily Carduelinae, serins, goldfinches, linnets, rose finches, grosbeaks, and allies. Check-list of the birds of the world, vol. 14. Cambridge: Museum of Comparative Zoology.
- Hudson R.R. 2002. Neutral model of genetic variation. *Bioinforma. Appl. Note*. 18:337–338.
- Hudson R.R., Boos D.D., Kaplan N.L. 1992. A statistical test for detecting geographic subdivision. *Mol. Biol. Evol.* 9:138–151.
- James P.M.A., Cooke B., Brunet B.M.T., Lumley L.M., Sperling F.A.H., Fortin M.J., Quinn V.S., Sturtevant B.R. 2015. Life-stage differences in spatial genetic structure in an irruptive forest insect: implications for dispersal and spatial synchrony. *Mol. Ecol.* 24:296–309.
- Johnson R.E. 1972. The biosystematics of the avian genus *Leucosticte*. Dissertation Submitted in Partial Satisfaction of the Requirements for the Degree of Doctor of Philosophy. Berkeley (CA): University of California Berkeley.
- Johnson R.E. 1975. New breeding localities for *Leucosticte* in the contiguous western United States. *The Auk*. 92:586–589.
- Johnson R.E. 2002. Black Rosy-Finch (*Leucosticte atrata*), version 2.0. In: Poole A.F., Gill F.B., editors. *The birds of North America*. Ithaca (NY): Cornell Lab of Ornithology.
- Keyghobadi N., Roland J., Strobeck C. 2005. Genetic differentiation and gene flow among populations of the alpine butterfly, *Parnassius smintheus*, vary with landscape connectivity. *Mol. Ecol.* 14:1897–1909.
- Li H., Durbin R. 2009. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. 26:589–595.
- Linck E., Battey C.J. 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic datasets. *Mol. Ecol. Resour.* 19:639–647.
- Lincoln F.C. 1916. The discovery of the nest and eggs of *Leucosticte australis*. *Auk*. 33:41–42.
- Llorens T.M., Tapper S.L., Coates D.J., McArthur S., Hankinson M., Byrne M. 2017. Does population distribution matter? Influence of a patchy versus continuous distribution on genetic patterns in a wind-pollinated shrub. *J. Biogeogr.* 44:361–374.
- Long C., Kubatko L. 2018. The effect of gene flow on coalescent-based species-tree inference. *Syst. Biol.* 67:770–785.
- Macdougall-Shackleton S.A., Johnson R.E., Hahn T.P. 2000. Gray-crowned Rosy-Finch (*Leucosticte tephrocotis*), version 2.0. In: Poole A.F., Gill F.B., editors. *The birds of North America*. Ithaca (NY): Cornell Lab of Ornithology.
- Maddison W.P. 1997. Gene Trees in Species Trees. *Syst. Biol.* 46:523–536.
- Marques D.A., Lucek K., Haesler M.P., Feller A.F., Meier J.I., Wagner C.E., Excoffier L., Seehausen O. 2017. Genomic landscape of early ecological speciation initiated by selection on nuptial colour. *Mol. Ecol.* 26:7–24.
- Marques D.A., Meier J.I., Seehausen O. 2019. A combinatorial view on speciation and adaptive radiation. *Trends Ecol. Evol.* 34:531–544.
- Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32:244–257.
- Mason N.A., Taylor S.A. 2015. Differentially expressed genes match bill morphology and plumage despite largely undifferentiated genomes in a Holarctic songbird. *Mol. Ecol.* 24:3009–3025.
- Mayr E. 1942. Systematics and the origin of species. New York (NY): Columbia University Press.
- McKenna A., Hanna M., Banks E., Sivachenko A.Y., Cibulskis K., Kernytsky A.M., Garimella K. V., Altshuler D., Gabriel S., Daly M., DePristo M.A. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequence data. *Genome Res.* 20:1297–1303.
- Miller H.C., Lambert D.M. 2004. Genetic drift outweighs balancing selection in shaping post-bottleneck major histocompatibility complex variation in New Zealand robins (Petroicidae). *Mol. Ecol.* 13:3709–3721.
- Morales A.E., Jackson N.D., Dewey T.A., O'Meara B.C., Carstens B.C. 2017. Speciation with gene flow in North American *Myotis* bats. *Syst. Biol.* 66:440–452.
- Nadeau N.J., Martin S.H., Kozak K.M., Salazar C., Dasmahapatra K.K., Davey J.W., Baxter S.W., Mark L. 2013. Genome-wide patterns of divergence and gene flow across a butterfly radiation. *Mol. Ecol.* 22:814–826.
- Neigel J.E., Avise J.C. 1986. Phylogenetic relationships of mitochondrial DNA under various demographic models of speciation. In: Karlin S., Nevo E., editors. *Evolutionary Processes and Theory*. Orlando (FL): Academic Press, INC. p. 515–534.
- Nosil P. 2008. Speciation with gene flow could be. *Mol. Ecol.* 17:2103–2106.
- Nychka D., Furrer R., Paige J., Sain S. 2017. fields: tools for spatial data. R package version 10.3. Available from: <https://github.com/NCAR/Fields>.
- Packard F.M. 1945. The birds of rocky mountain National Park. *Auk*. 62:371–394.
- Pavlidis P., Laurent S., Stephan W. 2010. MsABC: a modification of Hudson's ms to facilitate multi-locus ABC analysis. *Mol. Ecol. Resour.* 10:723–727.
- Pinho C., Hey J. 2010. Divergence with gene flow: models and data. *Annu. Rev. Ecol. Syst.* 41:215–230.
- Poplin R., Ruano-rubio V., DePristo M.A., Fennell T.J., Carneiro M.O., Auwera G.A. Van Der, Kling D.E., Gauthier D., Levy-moonshine A., Roazen D., Shakir K. 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *BioRxiv*. 1–22. doi: 10.1101/201178.
- Pruett C.L., Winker K. 2008. Evidence for cryptic northern refugia among high- and temperate-latitude species in Beringia: a response to Stewart and Dalén (2008). *Clim. Change*. 86:23–27.
- Pudlo P., Marin J.M., Estoup A., Cornuet J.M., Gautier M., Robert C.P. 2015. Reliable ABC model choice via random forests. *Bioinformatics*. 32:859–866.
- Rands C.M., Darling A., Fujita M., Kong L., Webster M.T., Clabaut C., Emes R.D., Heger A., Meader S., Hawkins M.B., Eisen M.B., Teiling C., Affourtit J., Boese B., Grant P.R., Grant B.R., Eisen J.A., Abzhanov A., Ponting C.P. 2013. Insights into the evolution of Darwin's finches from comparative analysis of the *Geospiza magnirostris* genome sequence. *BMC Genomics*. 14:1.
- Ravinet M., Faria R., Butlin R.K., Galindo J., Bierne N., Rafajlovič M., Noor M.A.F., Mehlig B., Westram A.M. 2017. Interpreting the genomic landscape of speciation: a road map for finding barriers to gene flow. *J. Evol. Biol.* 30:1450–1477.
- Richards C.M. 2000. Inbreeding depression and genetic rescue in a plant metapopulation. *Am. Nat.* 155:383–394.
- Sánchez-Montes G., Wang J., Ariño A.H., Martínez-Solano Í. 2018. Mountains as barriers to gene flow in amphibians: quantifying the differential effect of a major mountain ridge on the genetic structure

- of four sympatric species with different life history traits. *J. Biogeogr.* 45:318–331.
- Schwartz M.K., Luikart G., Waples R.S. 2007. Genetic monitoring as a promising tool for conservation and management. *Trends Ecol. Evol.* 22:25–33.
- Shultz A.J. 2017. Comparative and population genomics of host-pathogen co-evolution in birds. Dissertation Submitted in Partial Satisfaction of the Requirements for the Degree of Doctor of Philosophy. Cambridge (MA): Harvard University.
- Slatkin M. 1987. Gene flow and the geographic structure of natural populations. *Science.* 236:787–792.
- Smith C.C.R., Flaxman S.M., Scordato E.S.C., Kane N.C., Hund A.K., Sheta B.M., Safran R.J. 2018. Demographic inference in barn swallows using whole-genome data shows signal for bottleneck and subspecies differentiation during the Holocene. *Mol. Ecol.* 27:4200–4212.
- Solis-Lemus C., Ané C. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12:1–21.
- Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* 30:1312–1313.
- Starrett J., Derkarabetian S., Hedin M., Bryson R.W., McCormack J.E., Faircloth B.C. 2017. High phylogenetic utility of an ultraconserved element probe set designed for Arachnida. *Mol. Ecol. Resour.* 17:812–823.
- Sutton J.T., Nakagawa S., Robertson B.C., Jamieson I.G. 2011. Disentangling the roles of natural selection and genetic drift in shaping variation at MHC immunity genes. *Mol. Ecol.* 20:4408–4420.
- Swofford D.L. 2003. PAUP\*: phylogenetic analysis using parsimony (\*and other methods). Sunderland (MA): Sinauer Associates.
- Teitelbaum C.S., Mueller T. 2019. Beyond migration: causes and consequences of nomadic animal movements. *Trends Ecol. Evol.* 34:569–581.
- Toews D.P.L., Brelsford A., Gossen C., Milá B., Irwin D.E. 2016a. Genomic variation across the Yellow-rumped Warbler species complex. *Auk.* 133:698–717.
- Toews D.P.L., Taylor S.A., Vallender R., Brelsford A., Butcher B.G., Messer P.W., Lovette I.J. 2016b. Plumage genes and little else distinguish the genomes of hybridizing warblers. *Curr. Biol.* 26:2313–2318.
- Vaurie C. 1956. Systematic notes on palearctic birds. No. 20 Fringillidae: the Genera *Leucosticte*, *Rhodopechys*, *Carpodacus*, *Pinicola*, *Loxia*, *Uragus*, *Urocynchramus*, and *Propyrrhula*. *American Museum Novitates.* 1786:1–38.
- Winker A.K., Gibson D.D., SOWLS A.L., Lawhead B.E., Martin P.D., Hoberg E.P., Causey D. 2002. The birds of St. Matthew Island, Bering Sea. *Wilson Bull.* 114:491–509.
- Wolf J.B.W., Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. *Nat. Rev. Genet.* 18:87–100.
- Wright S. 1943. Isolation by distance. *Genetics.* 28:114–138.
- Wysoker A., Fennell T., Marth G., Abecasis G., Ruan J., Li H., Durbin R., Homer N., Handsaker B. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics.* 25:2078–2079.
- Zarza E., Faircloth B.C., Tsai W.L.E., Bryson R.W., Klicka J., McCormack J.E. 2016. Hidden histories of gene flow in highland birds revealed with genomic markers. *Mol. Ecol.* 25:5144–5157.
- Zheng X., Levine D., Shen J., Gogarten S.M., Laurie C., Weir B.S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics.* 28:3326–3328.