Low-coverage whole-genome sequencing for highly accurate population assignment: Mapping migratory connectivity in the American redstart (Setophaga ruticilla)

Matthew DeSaix[1], Eric Anderson[2], Christen Bossu[1], Christine Rayne[3], Teia Schweizer[1], Nicholas Bayly[4], Darshan Narang[5], Julie Hagelin[6], H. Lisle Gibbs[7], James Saracco[8], Thomas W Sherry[9], Michael Webster[10], Thomas Smith[11], Peter Marra[12], and Kristen Ruegg[1]

[1]Colorado State University
[2]NOAA Fisheries Southwest Fisheries Science Center Fisheries Ecology Division
[3]Colorado State University Department of Biology
[4]SELVA Investigación para la conservación en el Neotropico
[5]Trinidad and Tobago Field Naturalists' Club
[6]State of Alaska Department of Fish and Game
[7]The Ohio State University
[8]The Institute for Bird Populations
[9]Tulane University
[10]Cornell Lab of Ornithology
[11]UCLA Center for Tropical Research
[12]Georgetown University

June 16, 2023

Abstract

Understanding the geographic linkages among populations across the annual cycle is an essential component for understanding the ecology and evolution of migratory species and for facilitating their effective conservation. While genetic markers have been widely applied to describe migratory connections, the rapid development of new sequencing methods, such as low-coverage whole genome sequencing (lcWGS), provides new opportunities for improved estimates of migratory connectivity. Here, we use lcWGS to identify fine-scale population structure in a widespread songbird, the American Redstart (Setophaga ruticilla), and accurately assign individuals to genetically distinct breeding populations. Assignment of individuals from the nonbreeding range reveals population-specific patterns of varying migratory connectivity. By combining migratory connectivity results with demographic analysis of population abundance and trends, we consider full annual cycle conservation strategies for preserving numbers of individuals and genetic diversity. Notably, we highlight the importance of the Northern Temperate-Greater Antilles migratory population as containing the largest proportion of individuals in the species. Finally, we highlight valuable considerations for other population assignment studies aimed at using lcWGS. Our results have broad implications for improving our understanding of the ecology and evolution of migratory species through conservation genomics approaches.

Low-coverage whole genome sequencing for highly accurate population assignment: Mapping migratory connectivity in the American Redstart (*Setophaga ruticilla*)

Running title: American Redstart migratory connectivity

Authors : Matthew G. DeSaix[1], Eric C. Anderson[2,1,3], Christen M. Bossu[1], Christine E. Rayne[1], Teia M. Schweizer[3], Nicholas J. Bayly[4], Darshan S. Narang[5], Julie C. Hagelin[6], H. Lisle Gibbs[7,8], James F. Saracco[9], Thomas W. Sherry[10,11], Michael S. Webster[12,13], Thomas B. Smith[14,15], Peter P. Marra[16,17], Kristen C. Ruegg[1]

Affiliations :

[1] Department of Biology, Colorado State University, Fort Collins, Colorado, USA

[2] Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, Santa Cruz, California, USA

[3] Department of Fisheries, Wildlife, and Conservation Biology, Colorado State University, Fort Collins, Colorado, USA

[4] SELVA Investigación para la conservación en el Neotropico, DG42A #20-37, Bogotá, Colombia

[5] Trinidad and Tobago Field Naturalists' Club, Port of Spain, Trinidad and Tobago

[6] Threatened, Endangered and Diversity Program, Alaska Department of Fish and Game, Fairbanks, Alaska, USA

[7] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, Ohio, USA

[8] Ohio Biodiversity Conservation Partnership, The Ohio State University, Columbus, Ohio, USA

[9] The Institute for Bird Populations, Petaluma, California, USA

[10] Department of Ecology and Evolutionary Biology, Tulane University, New Orleans, Louisiana, USA

[11] Department of Biological Sciences, Dartmouth College, Hanover, New Hampshire, USA

[12] Cornell Lab of Ornithology, Ithaca, New York, USA

[13] Department of Neurobiology and Behavior, Cornell University, Ithaca, New York, USA

[14] Center for Tropical Research, Institute of the Environment & Sustainability, University of California Los Angeles, Los Angeles, California, USA

[15] Department of Ecology and Evolutionary Biology, University of California Los Angeles, Los Angeles, California, USA

[16] Department of Biology, Georgetown University, Washington, DC, USA

[17] McCourt School of Public Policy, Georgetown University, Washington, DC, USA

Corresponding author: mgdesaix@gmail.com

Abstract

Understanding the geographic linkages among populations across the annual cycle is an essential component for understanding the ecology and evolution of migratory species and for facilitating their effective conservation. While genetic markers have been widely applied to describe migratory connections, the rapid development of new sequencing methods, such as low-coverage whole genome sequencing (lcWGS), provides new opportunities for improved estimates of migratory connectivity. Here, we use lcWGS to identify fine-scale population structure in a widespread songbird, the American Redstart (*Setophaga ruticilla*), and accurately assign individuals to genetically distinct breeding populations. Assignment of individuals from the nonbreeding range reveals population-specific patterns of varying migratory connectivity. By combining migratory connectivity results with demographic analysis of population abundance and trends, we consider full annual cycle conservation strategies for preserving numbers of individuals and genetic diversity. Notably, we highlight the importance of the Northern Temperate-Greater Antilles migratory population as containing

the largest proportion of individuals in the species. Finally, we highlight valuable considerations for other population assignment studies aimed at using lcWGS. Our results have broad implications for improving our understanding of the ecology and evolution of migratory species through conservation genomics approaches.

Keywords : assignment bias, conservation genomics, effective sample size, migration, population genetics, songbird

Introduction

Long-distance migratory species pose distinct challenges to studies of ecology, evolution, and conservation because they occupy different geographical regions throughout the year that can be separated by thousands of kilometers. At each stage in the migratory annual cycle, migrant populations are subject to various stressors that can influence their fitness . As a result, effective conservation efforts require understanding migratory connectivity, defined as the links between different geographic regions used across the annual cycle . In the past 20 years, population genetics has become a well-established means for tracking migratory populations, especially for studies involving large sample sizes or small-bodied individuals . However, the value of genetic markers is often limited by the amount of genetic differentiation in a species and the availability of genetic data from individuals across the annual cycle .

Population assignment methods originated in the early 1980s and 1990s as a means of identifying breeding origins of migratory individuals back to distinct tributaries (in the case of fish) or geographic regions (in the case of bears) . Early methods relied on genetic markers that were limited to identifying only deep phylogeographic breaks within species . In recent years, next generation sequencing has facilitated the screening of a significantly larger number of genetic markers allowing for the delineation of breeding populations at finer spatial scales . Cost-effective delineation of patterns of migratory connectivity was made possible by designing single nucleotide polymorphisms (SNP) assays for a subset of these markers that were particularly useful for population assignment . While recent reductions in the cost of whole genome sequencing have made it possible to directly use low-coverage whole genome sequencing (lcWGS) data to screen migrant samples, the lack of software capable of dealing with the increase in marker number has prevented this method from being used for population assignment (DeSaix et al. in review).

Low-coverage WGS has made sequencing more affordable for non-model organisms by reducing the sequencing effort per individual, however it has distinct challenges. One of these challenges is dealing with low sequencing read depths per individual, which necessitates the use of probabilistic frameworks for genotype calling to account for the uncertainty inherent in the data . Accurate estimates of parameters such as allele frequency can be obtained by prioritizing larger sample sizes of individuals with lower sequencing depth . Guidelines for achieving accurate allele frequency estimation with lcWGS include sequencing individuals at a minimum of 1X coverage or having at least 10 individuals sequenced with a total sequencing depth of at least 10X . To take advantage of lcWGS data for population assignment, DeSaix et al. (*in review*) recently developed a software package, WGSassign, that accounts for uncertainty inherent to lcWGS data in population assignment tests. Here, for the first time, we use lcWGS data to assign migrants to their population of origin.

The American Redstart (*Setophaga ruticilla*) is an ideal system for evaluating the potential gains in effectiveness achievable by using lcWGS data for population assignment because previous studies using a variety of methods provide a strong foundation for comparisons. The American Redstart is a widely distributed migratory songbird with a breeding distribution across North America and stationary nonbreeding distribution throughout the Caribbean, northern South America, Central America, and Mexico . For several decades, the American Redstart has been a model species for understanding migratory ecology and has been used to elucidate territoriality on the wintering grounds , foraging behavior , habitat selection , and carry-over effects of stressors across the annual cycle . Phylogeographic structure has previously been detected between a small region in the Maritime Provinces, specifically in Newfoundland and New Brunswick in the northeastern portion of the range, and the rest of the continental breeding range using mtDNA . Subsequent analysis of migratory connectivity using mtDNA revealed that Newfoundland breeders overwintered on the

islands of Puerto Rico and the Dominican Republic, while continental breeding birds overwintered across the entire nonbreeding range . Stable isotope studies have shown strong migratory connectivity, with eastern breeding birds overwintering in the Caribbean and western breeding birds overwintering in Central America and Mexico , but whether these migratory differences correspond to genetic differentiation has not been tested.

Here we aim to demonstrate the effectiveness of using lcWGS data for population assignment of nonbreeding individual using the American Redstart as a model species. Our main objectives were: 1) Identify population-specific migratory connectivity in the American Redstart using lcWGS data, 2) Assess conservation implications of migratory connectivity by identifying relative abundance and trends in population size, and 3) Provide study design recommendations to facilitate the use of lcWGS data in other population assignment studies. Our results have broad implications for improving our understanding of the ecology and evolution of migratory species through conservation genomics approaches.

Methods

Genetic sampling and library preparation

Sample site locations were chosen to maximize sampling coverage across the breeding and nonbreeding ranges of the American Redstart. We used genetic samples from a total of 330 individuals: 182 individuals from 16 locations across the breeding range and 148 individuals from 15 locations in the nonbreeding range (Table S1). Sample collection occurred between 1993 and 2022 and consisted of either blood from brachial venipuncture or feathers. We extracted DNA from blood samples using the standard protocol for Qiagen DNEasy Blood and Tissue Kits and we modified the protocol to maximize DNA yield from feathers. Whole genome sequencing libraries were prepared following modifications of Illumina's Nextera Library Preparation protocol . Pooled libraries were sequenced on eight HiSeq 4000 lanes at Novogene Corporation Inc with a target sequencing depth of 2X per individual.

Bioinformatics

We trimmed the sequence data to remove potential PCR artifacts using the program TrimGalore version 0.6.5 (https://github.com/FelixKrueger/TrimGalore), a wrapper for Cutadapt . We used the Burrows-Wheeler Aligner software version 0.7.17 to map reads to a reference genome from the closely related Yellow Warbler (*Setophaga petechia* ; Bay et al. 2018). After mapping, the resulting SAM files were sorted, converted to BAM files, and indexed using Samtools version 1.9 . We marked read duplicates with MarkDuplicates from GATK version 4.1.4.0 and clipped overlapping reads with the clipOverlap function from bamUtil (https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap). Sequencing depth for individuals was calculated using Samtools. Initial population genetics analyses revealed a large effect in the data due to high variation in sequencing depth among individuals. To reduce sequencing depth variation, we followed the recommendations of and used the DownsampleSam function from GATK to randomly down sample reads from BAM files with greater than 2X coverage, to 2X coverage.

To identify genetic markers from low-coverage WGS data, we used stringent filtering options in ANGSD version 0.9.40 (). We retained reads with a mapping quality of at least 30 and base quality of at least 33. SNPs were identified based on a p-value of less than 1e-6. We retained SNPs that had read data in at least 50% of individuals ($n = 165$), a minor allele frequency greater than 0.05, and minimum and maximum total depths of 231 and 924, respectively. The minimum total depth threshold was chosen by the minimum number of individuals required to call a variant ($n = 165$) multiplied by the mean sequencing depth of all individuals (1.4X). The maximum total depth threshold was determined by 2 * total number of individuals * mean sequencing depth. The filtered variants were output as genotype likelihoods and used in subsequent analyses.

Genetically distinct breeding populations

Given that signatures of population structure can be skewed by closely related individuals, we used NGSrelate version 2 to identify and remove individuals with with up to second-degree relationships (kinship > 0.0884).

We implemented principal components analysis (PCA) and estimated individual admixture proportions using Pcangsd , which estimates individual allele frequencies to minimize bias from low and variable sequencing depth. We determined the number of genetically distinct breeding populations of American Redstarts by identifying congruent geographic signatures of clustering in the PCA with groupings of individuals based on admixture proportions. Posterior probabilities of group membership from the admixture proportions were visualized on a base map from Natural Earth (naturalearthdata.com) with each group specified by a different colour, and clipped to the breeding range of the American Redstart (Strimas-Mackey et al., 2021). Colour transparency was scaled such that the highest posterior probability of group membership is opaque while the smallest posterior probability is transparent. Visualization was performed in R (R Core Team, 2021). Genetic differentiation among the breeding populations was calculated by creating site allele frequency files for each breeding population and calculating F_{ST} in ANGSD ().

Effective sample size and population assignment

To assess our ability to accurately assign individuals of unknown origin to breeding populations, we first determined the accuracy of assignment of the known breeding origin individuals using a leave-one-out approach implemented in WGSassign (DeSaix et al. *in review*). Leave-one-out avoids assignment bias by iteratively removing an individual from their given source population, re-estimating the allele frequency of the source population, and then calculating the likelihood of the individual's assignment to each population. Another source of bias in assignment tests is variation in the precision of allele frequency estimation, which arises from populations having different numbers of samples and/or having differences in sequencing depth of their individuals. To mitigate this bias, we tested two other approaches for source population sampling design: 1) we reduced the number of samples per breeding population to be the same as the population with the fewest samples (size-standardized breeding populations; SSBPs) and 2) we followed the guidelines in DeSaix et al. (*in review*) to standardize the effective sample sizes of the breeding populations (effective-size-standardized breeding populations; ESSBPs).*Effective sample size* is a Fisher information metric that determines the comparable number of individuals with known genotypes that would reflect the same variance in estimated allele frequency as the sampled low-coverage individuals (DeSaix et al., *in review*). The purpose of ESSBPs is to equalize the effective sample size among populations by removing individuals from the populations with the highest effect sample sizes, thereby making the precision of allele frequency estimation similar among the different populations. We used WGSassign to calculate each breeding population's effective sample size for the SSBPs and ESSBPs and performed leave-one-out assignment. We also performed standard assignment with all breeding individuals not in the standardized sets. Leave-one-out assignment for the full data set and the combined leave-one-out assignment and standard assignment accuracy were compared across all three source population sampling designs. Posterior probabilities of assignment to a population were determined by dividing the maximum likelihood of assignment over the sum of all likelihoods. A cut-off of 0.8 was used for the posterior probability to determine if an individual was confidently assigned to a population.

Low-coverage and population assignment

Since the majority of our nonbreeding and breeding samples were feathers and blood, respectively, we expected the nonbreeding samples to have lower sequencing depth than breeding samples. Therefore, to ensure that we could still achieve high assignment accuracy at lower depths for the nonbreeding samples, which have unknown breeding origin, we first tested assignment accuracy with low coverage breeding samples of known origin. We used the set of individuals from our ESSBPs to estimate population allele frequencies (our training set) and used the remaining breeding samples as a test set. We created two data sets from the test set individuals by down sampling to 0.1X and 0.01X. These two thresholds were based on the majority of the nonbreeding samples being greater than 0.1X and the lowest coverage sample being 0.02X. To determine the accuracy of assignment of individuals with low sequencing depths, we assigned the test sets back to the standardized breeding populations and compared the population assigned with the known population of origin.

Determining breeding origin of individuals on the nonbreeding range

We assigned individuals sampled from the nonbreeding range to the ESSBPs using WGSassign. Since these individuals are of unknown origin, we assumed the accuracy of their assignment would be comparable to the accuracy achieved with known breeding samples. Individuals at the periphery or boundaries of genetically distinct populations may have admixed genomes that are not truly representative of either population. While posterior probabilities of assignment are typically used to detect admixture and determine confidence of assignment, our preliminary results showed that posterior probabilities of assignment were unreliable with lcWGS data (see *Results* and *Discussion*). Therefore, we split up the genotype likelihood data into 10 subsets of 400,000 SNPs, in order of genomic region, for each individual and used a consistency of assignment threshold of 0.8 (*i.e.* , at least 8 of the 10 datasets being assigned to the same population) to determine confidence in assignment. We validated this approach on the 47 breeding individuals used as the testing set for the ESSBPs and then used it for the 148 individuals from the nonbreeding range.

Demographic analysis

We estimated relative population size indices and population trends (1968-2021) for each of the five breeding populations and across the entire breeding range using Breeding Bird Survey (BBS) data (Pardieck et al., 2020). We used a hierarchical over-dispersed Poisson model (Sauer et al., 2011) to analyze the BBS data. All BBS routes within a 50-km buffer of a polygon defining one of the breeding populations were assigned to that breeding population. This breeding population assignment was then included as the fixed stratum intercept and trend effects of the log-linear model of the Poisson mean. We estimated current (2017-2021) population size indices by summarizing posterior distributions of estimated mean route-level counts that were weighted by geographic area encompassed by the breeding population polygon and the proportion of routes in the polygon with American Redstart detections (Sauer & Link, 2011). Long-term trends in population size were estimated as the geometric mean of yearly changes from 1968-2021 (Sauer & Link, 2011). We implemented the hierarchical model in JAGS 4.3.1 (Plummer, 2003) using the jagsUI (Kellner & Meredith, 2021) package in R (R Core Team, 2022). We assigned vague prior distributions for all model parameters and hyperparameters. Posterior distributions were derived from 40,000 simulated values of four chains from the posterior distribution after an adaptive phase of 20,000 iterations and burn-in of 10,000 samples of the Gibbs sampler and thinning by 3. Markov chains were determined to have successfully converged based on $\hat{R}$<1.1 for posterior estimates of all parameters (Gelman & Hill, 2007).

Results

Genetically distinct breeding populations

Sequencing efforts resulted in sequences from 330 individuals with a mean coverage of 1.6X (range: 0.02X – 5.2X; Table S1). For the 182 breeding samples, the mean coverage was 1.7X (range: 0.6X – 5.2X), while the 148 nonbreeding samples had a mean coverage of 1.5X (range: 0.02 – 3.4X). Down sampling individuals above 2X coverage to 2X coverage, resulted in an overall mean coverage of 1.4X (1.5X for breeding samples, 1.3X for nonbreeding samples). Our SNP filtering produced genotype likelihood data for 4,722,390 variants. We removed 13 individuals from subsequent analyses due to high relatedness by removing a single individual from each related pair. Principal components analysis with the breeding samples revealed five genetic clusters that aligned with geography: Western Boreal (Alaska to Saskatchewan), Basin Rockies (South Dakota and Montana), Southern Temperate (from Missouri, east to Maryland, south to Louisiana), Northern Temperate (from Minnesota, east to Quebec, south to Pennsylvania), and Maritime Provinces (New Brunswick and Newfoundland; Figure S1). Admixture results for five groups revealed a similar delineation of individuals as in the principal components analysis (Figure 1, Figure S1). Pairwise F_{ST} values among these breeding populations had a mean of 0.009 and ranged from the weakest differentiation (F_{ST} = 0.004) between the Northern Temperate and Southern Temperate groups and the strongest (F_{ST} = 0.018) between Maritime Provinces and Basin Rockies (Table S2). Based on our comprehensive sampling of the core regions of the American Redstart breeding range and the population structure results, we expect there to be no unsampled genetically distinct breeding populations. For example, while we did not sample individuals from Alaska, we expect Alaska breeding birds to be a part of the Western Boreal breeding population.

Effective sample size and assignment

Breeding populations ranged in number of samples from 27 (Maritime Provinces and Basin Rockies) to 47 (Southern Temperate) and ranged in effective sample size from 12.3 – 24.6. Mean accuracy of leave-one-out assignment with these individuals was 89.3% (151 out of 169 individuals), and accuracy by breeding population ranged from 63.0% (Maritime Provinces) to 100% (Southern Temperate and Basin Rockies). All 18 individuals that were inaccurately assigned were assigned to a breeding population with higher effective sample size than the known breeding population (Figure 2). All posterior probabilities of assignment (accurate and inaccurate) were greater than 0.8 (Table S3).

The SSBPs all had 27 samples but ranged in effective sample size from 12.3 – 16.1. The mean accuracy of assignment was 97.0% (164 out of 169 individuals) for the leave-one-out assignment with the 135 individuals in the SSBPs and standard assignment for the remaining 34 individuals. Three individuals were incorrectly assigned in the leave-one-out assignment test and two individuals were incorrectly assigned in the standard assignment test. All five incorrectly assigned individuals were assigned to a breeding population with higher effective sample size than the known breeding population (Figure 2).

The ESSBPs ranged in number of samples from 21 (Basin Rockies) to 27 (Maritime Provinces) but had minimal variation in their effective sample sizes (range: 12.0 – 12.5). Mean assignment accuracy was 99.4% (168 out of 169 individuals) for the leave-one-out-assignment with the 122 individuals in the ESSBPs and standard assignment for the remaining 47 individuals. Only one individual was incorrectly assigned from the Northern Temperate population to the Southern Temperate population, and this did not correspond to a breeding population with higher effective sample size. Interestingly, this same individual (sampled in Minnesota from the Northern Temperate population) also stands out in the PCA results as clustering more closely with individuals from the Southern Temperate population (Figure S1). Given the higher accuracy of assignment with the ESSBPs, compared to the other sets of breeding individuals, we continued subsequent assignment testing using only the ESSBPs data set as the source populations.

Testing sequence depth thresholds for assignment

To test whether lower coverage data would affect our ability to accurately assign breeding individuals, we used the 47 individuals from the breeding populations, that were not used in the ESSBPs, as a testing set for further down sampling. We did not down sample individuals from the ESSBPs because we did not want to lower the effective sample size of the source populations, but rather test how well we could assign individuals with lower coverage given the effective sample sizes of our source populations (and the amount of genetic differentiation among them). The testing set consisted of 22 individuals from the Southern Temperate population, 14 individuals from the Northern Temperate population, 6 individuals from Basin Rockies population, and 5 individuals from the Western Boreal population. The sequencing depth of these 47 individuals ranged from 0.6X – 2.0X. Down sampling these individuals to 0.1X resulted in 100% assignment accuracy, and further down sampling to 0.01X resulted in 97.9% accuracy (one individual from the Southern Temperate population assigned to the Northern Temperate population; Table 1, Table S4). The individual incorrectly assigned from the Southern Temperate population was from a sampling site in Pennsylvania which is on the border of our boundary for the Southern Temperate and Northern Temperate populations.

Nonbreeding assignment

Implicit in the assignment of the nonbreeding individuals to breeding populations is that the breeding origin of these individuals is unknown. Given that assignment to the ESSBPs had an accuracy of 99.4% (168 out of 169 samples) for individuals with sequencing coverage of 0.6X – 2.0X, and accuracy of 97.9% (46 out of 47 individuals) for individuals down sampled to sequencing coverage of 0.01X, we assumed that we could correctly assign nonbreeding individuals (sequencing coverage range: 0.02X – 2.0X, mean 1.3X) with high confidence. Assignment of the 148 nonbreeding individuals resulted in the largest number of individuals being assigned to the Northern Temperate population (n = 64) and the least number of individuals being assigned to the Basin Rockies population (n = 2; Table S5). Of the 148 individuals, 139 individuals had assignment consistency of at least 0.8 for the 10 subsets of data, and these individuals were used to infer migratory

connectivity. Testing consistency of assignment on the 47 breeding individuals identified three individuals with assignment consistency of < 0.8. One of these individuals from Minnesota was previously identified as an outlier in the PCA, and the other two individuals were from Pennsylvania, which is on the boundary of the Southern Temperate and Northern Temperate populations.

Mapping of the nonbreeding assignment results revealed patterns of strong migratory connectivity across the breeding range. Notably, the Maritime Provinces breeding population has strong connectivity with eastern Colombia, the Northern Temperate breeding population with the Greater Antilles, the Southern Temperate breeding population with the Lesser Antilles, and the Western Boreal breeding population with Central America and Mexico.

Demographic analysis

Using 1,766 BBS routes from 1968-2021, we estimated the range-wide trend in population size to be -0.29% per year (95% CI: -0.57, -0.02). Trends among the breeding populations were variable (Table 2). The Northern Temperate breeding population was estimated to be increasing by 0.67% per year (95% CI: 0.33, 1.01). The Southern Temperate breeding population was estimated to be declining by 0.34% per year (95% CI: -0.75, 0.06), but the credible intervals were overlapping 0, thus indicating potential stability in that population. The remaining three populations (Basin Rockies, Maritime Provinces, and Western Boreal) were all estimated to be declining and had negative values for the upper bounds of the 95% credible intervals. The Northern Temperate population had the highest relative abundance of 3.70 (95% CI: 3.09, 4.52), followed by Maritime Provinces (1.96; 95% CI: 1.57, 2.49), Western Boreal (0.66; 95% CI: 0.52, 0.84), Southern Temperate (0.15; 95% CI: 0.13, 0.17), and Basin Rockies (0.01; 95% CI: 0.01, 0.02). The density of the number of birds per BBS route was highest in the Maritime Provinces (30.56; 95% CI: 24.33, 38.60), followed by Northern Temperate (15.16; 95% CI: 12.68, 18.54), Western Boreal (1.66; 95% CI: 1.31, 2.10), Southern Temperate (0.67; 95% CI: 0.58, 0.78), and Basin Rockies (0.16; 95% CI: 0.10, 0.25).

Discussion

The results of this study demonstrate that lcWGS data is well-suited for highly accurate population assignment, even with weakly differentiated population structure. In the American Redstart, lcWGS data provided an improvement over previous migratory connectivity studies using genetic and stable isotope data by allowing us to identify five genetically distinct breeding populations and clearly delineate population-specific nonbreeding ranges. Identifying migratory connectivity of genetically distinct populations is an essential step toward full annual cycle conservation aimed at preserving unique genetic variation. To this end, we integrate the migratory connectivity results with analysis of population abundance and trends to demonstrate the conservation implications of the observed population-specific migratory patterns. More broadly, we also show that when using lcWGS data for population assignment it is essential to implement a sampling design that balances*effective sample size* across source populations to avoid assignment bias that arises from variation in sequencing depth and population sample size.

Mapping migratory connectivity

Population structure analyses identified five genetically distinct breeding populations with weak genetic differentiation, in contrast to a previous mtDNA analysis that identified only two populations split by a phylogeographic break (Colbeck et al., 2008). Our delineation of the Maritime Provinces breeding population in the far northeast portion of the range corresponds with the Newfoundland population from the mtDNA analysis. Colbeck et al. (2008) hypothesized that the phylogeographic separation of Newfoundland and mainland American Redstart populations was the result of two refugia during Pleistocene glaciations. Our findings of weak genetic differentiation between the Maritime Provinces and other breeding populations suggest that there is ongoing gene flow among these populations. However, the limited admixture of individuals sampled in Newfoundland supports the notion that geographic separation of the island provides some barrier to gene flow, which has been demonstrated in several other avian species . The weakest genetic differentiation was found among the Western Boreal, Northern Temperate, and Southern Temperate breeding populations (F_{ST}: 0.004-0.006), suggesting limited barriers to gene flow. The Basin Rockies breeding population had

higher genetic differentiation with the eastern breeding populations than the more northern Western Boreal population, which corresponds to the Great Plains functioning as a barrier to gene flow.

Using the five genetically distinct breeding populations allowed us to document at a fine scale more complex migratory patterns than previously identified. At the continental scale, our results broadly correspond to previous stable isotope studies that found eastern breeding American Redstarts overwintered in the eastern nonbreeding range and western breeders overwintered in the west . In several other species of Nearctic–Neotropical migrants, similar patterns of parallel migration have also been observed . However, in contrast to previous isotope analyses in American Redstart , our use of genomic data allowed us to clearly differentiate Maritime Provinces and Northern Temperate breeding birds and revealed that individuals breeding in the Maritime Provinces do not follow the parallel migration pattern. Parallel migration would result in these breeders being found in the far eastern portion of the nonbreeding range (e.g., Lesser Antilles and Trinidad and Tobago). Instead, individuals from the Maritime provinces bypass the Caribbean portion of the nonbreeding range and have a "leap-frog" migratory pattern to eastern Colombia. One explanation for the discordance of the Maritime Provinces migratory connectivity patterns from the rest of the breeding populations is the phylogeographic separation of these regions documented by Colbeck et al. (2008). Migration routes are influenced by the historical separation of Pleistocene glacial refugia and in the American Redstart, an Atlantic Shelf (near the Maritime Provinces) and eastern continental refugia are hypothesized to have caused the observed phylogeographic separation of these regions . A previous mtDNA analysis of American Redstart migratory connectivity only detected several individuals from the Maritime Provinces population in the Caribbean islands of the Dominican Republic and Puerto Rico but lacked samples from South America . Our results suggest that the Maritime Provinces breeding population has the strongest connectivity with eastern Colombia, though the full extent of the population's nonbreeding distribution throughout South America is unknown.

Our use of genomic data allowed us to characterize migratory connectivity at a fine scale and identify distinct regions on the wintering range that separate breeding populations. In southern Central America, a clear split between the two sampling sites in Costa Rica, a separation of 360 km, occurs where the northern site predominantly has individuals from the Western Boreal breeding population and the southern site has individuals from the Southern Temperate population. This split corresponds with a biogeographic separation of drier broadleaf forest in the northern Pacific side of Costa Rica and moist broadleaf forest in the southern Pacific side (Corrales, Bouroncle, & Zamora 2015). Another geographic split in breeding origin occurs between the Lesser Antilles (Southern Temperate) and the Greater Antilles (Northern Temperate) in the Caribbean. Sampling of American Redstarts in Colombia was limited to the eastern slopes of the East Andes, and may not represent the wider Andes, given that the three chains of the Andes that run through Colombia influence connectivity patterns in the Canada Warbler,*Cardellina canadensis* Further population assignment studies that include sampling of American Redstarts from the Central and Western Andes, and the Caribbean region of Colombia, may identify the Andes Mountains as another barrier in the nonbreeding region, creating geographic splits in breeding origin for this species.

Conservation implications

Describing migratory connectivity is essential for informing effective wildlife conservation and management decisions involving migratory species Our delineation of five breeding populations, and their linkages to wintering regions, provides the necessary information to prioritize regions for conservation and improve our understanding of the underlying drivers of abundance. While our demographic analysis highlights that the species is declining overall from 1966 – 2021, there is wide variation in trends among the breeding populations. The Northern Temperate population has the largest population of American Redstarts on the breeding grounds and is increasing in abundance. One potential explanation for the increase in abundance is that birds in the southern portion of the breeding range have shifted their breeding latitude northward in response to climate change, as has been documented in other Nearctic-Neotropical migrants . However, our results do not depict a correspondingly large decline in the Southern Temperate breeding population which would be the source population of northward movement. Given that the Northern Temperate breeding

population has strong connectivity with the Greater Antilles archipelago in the Caribbean, efforts aimed at conserving the greatest proportion of the global distribution of American Redstarts could focus on the Northern Temperate-Greater Antilles migratory population.

The Maritime Provinces population had the second highest abundance of American Redstarts and the highest density of individuals. Despite being geographically adjacent to the Northern Temperate breeding population, Maritime Provinces individuals were detected almost exclusively outside the Caribbean, along the eastern slopes of the Andes of Colombia. Our demographic analysis highlighted the Maritime Provinces to be the second fastest declining population. Thus, future research into the stressors driving this decline could focus on the breeding region as well as stationary nonbreeding region of eastern Colombia. Notably, other populations of long-distance migratory birds connected to the Eastern Andes are also experiencing declines, including populations of Canada Warbler (Wilson et al., 2018) and Cerulean Warbler, *Setophaga cerulea* (Raybuck et al., 2022). Additionally, in species such as the Canada Warbler, migration routes between North and South America can concentrate in small regions of Central America which can also affect population trends . Given the phylogeographic split of the Maritime Provinces breeding population with the mainland (Colbeck et al., 2008), conservation of this migratory population may also be important for preserving genetic diversity within the species. The Western Boreal population, ranging from Alaska to Saskatchewan, was characterized by the demographic analysis as having the third highest abundance and density, with population declines larger than the range-wide decline. Strong migratory connectivity with Mexico and Central America highlights the need for conservation efforts to focus on the most western portion of the range for this migratory population.

The Southern Temperate breeding population is unique in American Redstarts, in that nonbreeding individuals were sampled in both the far eastern Caribbean as well as in Central America. Our lack of sampling between these regions in northern South America precludes our ability to describe whether there is a migratory divide within the Southern Temperate population, or individuals are spread across this portion of the nonbreeding range. While weak connectivity across a large nonbreeding distribution could promote resilience from stressors on any single portion of the nonbreeding distribution , this makes targeting regions for conservation difficult.

Low-coverage WGS for population assignment

In addition to elucidating fine-scale migratory connectivity patterns in the American Redstart, our results provide important considerations for other population assignment studies using lcWGS. We found that balancing*effective sample sizes* of the source populations to within one effective individual of each other was essential for accurate assignment. Even when the actual number of individuals used per population was the same, variation in mean depth (1.3X – 1.9X) between populations skewed the effective sample sizes, resulting in decreased assignment accuracy. Other studies with known genotypes from RADseq have demonstrated the influence of actual sample size on overall assignment accuracy but not how it affects assignment bias . The effective sample sizes needed per population for accurate assignment and the degree of standardizing these values will depend on the population structure of the study system. For example, study systems with higher genetic differentiation between populations may not need to finely standardize effective sample size to achieve high assignment accuracy. We suggest that other population assignment studies similarly evaluate the influence of source population effective sample size on known source individuals before assigning individuals of unknown origin. Reducing the effective sample size of a sampled population can be achieved by either removing individuals or down sampling the read depth. In this study, we chose to remove individuals, and used the individuals' effective sample sizes as a guide for how many individuals to remove from each population (resulting in 21 – 27 samples per population). For studies with smaller sample sizes, it may be worthwhile to investigate if retaining all individuals, but down sampling reads is a better alternative for standardizing effective sample sizes to retain more variation from individuals.

Importantly, here we demonstrate that individuals with very low whole genome coverage (0.01X – 0.1X) can still be accurately assigned to source populations with sufficient effective sample sizes. These results suggest that increasing the number of samples and decreasing individual sequencing depth is an effective study design strategy for population assignment. For migratory connectivity studies, increased sampling (both number of

individuals at each location and the number of locations sampled) across nonbreeding stages of the annual cycle can drastically improve our understanding of population-level connectivity at low cost. Combined with cost-effective approaches for library preparation (e.g.), lcWGS is increasingly becoming economically feasible for a wide-range of studies. However, a trade-off with lcWGS is that the sequence data processing requires additional costs associated with time spent on the bioinformatics analysis. For studies interested in population assignment with a large number of samples, increasing the number of samples per lane, thereby decreasing the mean average sequencing depth, may make lcWGS economically feasible compared to other sequencing methods. For a comprehensive review of coverage guidelines for different types of analyses with low-coverage WGS data see Lou et al. (2021).

An interesting aspect of our results was that all posterior probabilities of assignment were > 0.8, even for potentially admixed individuals. A standard method to determine assignment confidence in population assignment studies is to use a cutoff value for posterior probabilities of assignment . Individuals with low posterior probabilities of assignment (e.g., < 0.8) can be highly admixed. Thus, it is inaccurate to classify them as from a specific population. However, we suspect that with lcWGS data, the high prevalence of loci with single read results in the likelihood being highest for a homozygous genotype. Thus, admixed individuals may "switch" their population of maximum likelihood depending on the loci used for assignment. Our use of an assignment consistency threshold addressed this concern by creating subsets of genomic data for population assignment to determine if individuals could reliably be assigned to a single population when different loci were used. Testing the assignment consistency threshold with known source individuals revealed three individuals with inconsistent assignment (< 0.8, i.e., 8 out of 10 genomic datasets) and were likely admixed between pure Northern Temperate and Southern Temperate populations. These results highlight that the consistency of assignment may be more reliable than posterior probabilities for confidently assigning individuals of unknown origin. Further development of spatially explicit assignment methods for genotype likelihood data would be helpful for determining the likely origin of admixed individuals at the periphery of source populations.

Conclusion

Low-coverage WGS is a powerful and potentially cost-effective approach for population assignment studies. We demonstrate that high assignment accuracy can be obtained for weakly differentiated populations, even for individuals with very low sequencing coverage (< 0.1X). We further demonstrate the importance of balancing the effective sample sizes of source populations to avoid assignment bias due to variation in the precision of allele frequency estimation. By applying these methods to the American Redstart, we reveal broad-scale parallel migration and highlight unique population-specific patterns of connectivity. In combination with our demographic analysis, we demonstrate the importance of the Northern Temperate-Greater Antilles migratory population to the total abundance of the species. Furthermore, our identification of nonbreeding regions for the genetically distinct breeding populations provides a foundation for a full annual cycle approach towards preserving genetic diversity. Together, our results provide a valuable framework for studies that aim to use lcWGS to understand the ecology and evolution of migratory species.

Acknowledgments

This study was funded through a State Wildlife Grant (SWG T-36, Project 2) administered by the Alaska Department of Fish and Game's Threatened, Endangered and Diversity Program, an NSF CAREER award (008933-00002), a generous donation from an anonymous donor, and a National Geographic grant (WW-202R-170) to KCR, as well as an NSF LTREB (No. 0649679, 1242584) grant to PPM and TWS, and funding from the Smithsonian Institute's James Bond Endowment Fund to PPM. Writing of this paper was initiated while MGD and ECA were scientists-in-residence in the Mobile High Altitude Venue for Ecological Analysis, Genetics, and Statistics on location in Moab, Utah for five days in March 2023. This is contribution number mHAVEAGAS-002. We gratefully acknowledge the services and kind staff at the Grand County Public Library in Moab. Field work was conducted under permits from the USGS, the Jamaican National Environment and Protection Agency, and Smithsonian National Zoo IACUC approval 14-03. Sample collection in Colombia was undertaken as part of a study funded by Environment and Climate Change Canada and

both sample collection and export were conducted under permits (resolución 00874) from the Autoridad Nacional de Licencias Ambientales (ANLA). Sample collection and export in Trinidad and Tobago were conducted under the Special Game License (Chapter 67:01 Section 10) and Special Export License from the Wildlife Section Forestry Division, and we thank Carl Fitzjames, Richard Smith, Richelle Smith, Shivam Mahadeo, and Laura Baboolal for sample collection. We thank Keith Hobson, Anne-Marie Barber, Lorie Collins, Robert Dawson, Lilie DeSousa, Jose Diaz, Emmanuel Milot, David Okines, and John Woods for sample collection and preparation and support from the Max Bell Foundation and Environment Canada. We thank staff of the Institute for Bird Populations and MAPS and MoSI program cooperators for providing samples or assisting with sample collection. We thank Jeanie Woltz, Junior Tremblay, Jacques Ibarzabal, Tim Kita, John Woodcock, Alexis Cerezo, Susan Koenig, Ingrid Tello-Lopez, Rafael Rueda Hernandez, Fred Schaffner, Stacey Hayden, Lori Walewski, Chantal Villeneuve, Chase Mendenhall, and David Curiel for sample collection. We thank Diana Baetscher for providing feedback on a draft of this manuscript. This work utilized the Alpine high performance computing resource at the University of Colorado Boulder. Alpine is jointly funded by the University of Colorado Boulder, the University of Colorado Anschutz, Colorado State University, and the National Science Foundation (award 2201538).

Conflict of interest statement

The authors declare no conflict of interest.

Data availability statement

Genomic data have been deposited in the Dryad repository [PUT DOI HERE FROM MOLECULAR ECOLOGY]. All analysis scripts have been made available at https://github.com/mgdesaix/amre-mc.

Benefit-sharing statement

An international research collaboration was developed with scientists facilitating genetic sample collection and all collaborators are included as co-authors. The contributions of all individuals to the research are described in the acknowledgments. The results of the research have been shared with the broader scientific community and address a priority concern of the conservation of migratory species.

Author Contributions

MGD and KCR designed the study. Sample collection was facilitated and performed by MGD, NJB, DSN, JCH, HLG, TWS, MSW, TBS, PPM, and KCR. Laboratory work was performed by CER and TMS. MGD performed the bioinformatics data generation. CMB provided bioinformatics support and data management. ECA contributed to the population assignment study design. MGD and JFS performed the data analysis. MGD and CMB generated the figures of the manuscript. MGD and KCR wrote the paper and all authors edited and provided feedback to manuscript drafts.

References

Figures

Figure 1. The population structure on the breeding range is delineated by five genetically distinct clusters (colored polygons) from the results of the admixture analysis (top panel). Population structure was determined using 169 individuals from 16 sites (black points) on the breeding range. Individuals sampled from the non-breeding range were determined to originate from a given breeding population through population assignment tests (n = 138; colored circles). The point colors on the nonbreeding range represent the breeding population of maximum likelihood of assignment and the extent of the nonbreeding range is provided by the grey polygon. Strong migratory connectivity is evident from the general separation of breeding population assignment across the wintering range.

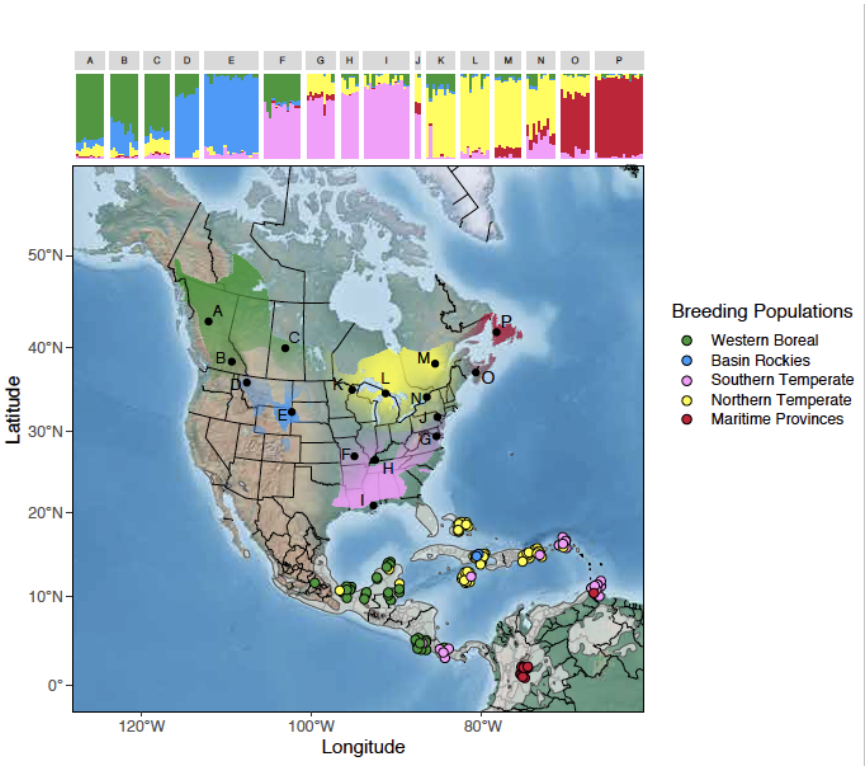


Figure 2. Population assignment of known breeding individuals revealed assignment bias from unequal effective sample sizes. Circles represent the breeding populations (colored), with circle size representing effective sample size, and arrows represent the assignment of individuals from their known breeding population to their assigned population. Arrows are scaled in size by the number of individuals assigned. Colored arrows represent the correct individuals assigned to a breeding population, whereas black arrows indicate incorrect assignment to a different population. A) When using all samples to calculate allele frequencies in breeding populations, all incorrectly assigned individuals (n = 18) were assigned to a population with higher effective sample size. B) Standardizing breeding populations by sample sizes (27 individuals per population) resulted in less incorrect assignment (n = 5), but all individuals were still assigned to another population with higher effective sample size. C) Standardizing breeding populations to approximately the same effective sample size (˜12 effective individuals), resulted in only one individual being incorrectly assigned. In all cases, incorrect assignment was typically to a geographically neighboring population.

Hosted file

`image2.emf` available at https://authorea.com/users/624774/articles/649788-low-coverage-whole-genome-sequencing-for-highly-accurate-population-assignment-mapping-migratory-connectivity-in-the-american-redstart-setophaga-ruticilla

Tables

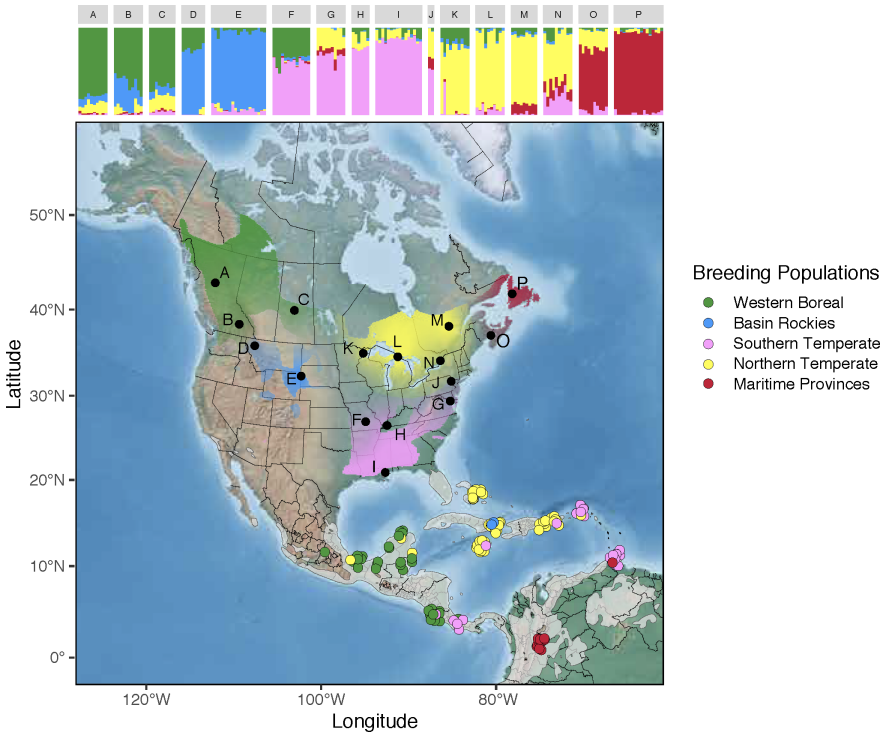
Table 1. Assignment accuracy of the known breeding samples used as the testing set (n = 47) for the effective size standardized breeding populations. Down sampling to both 0.1X and 0.01X achieved high assignment accuracy.

Depth	Western Boreal	Basin Rockies	Southern Temperate	Northern Temperate	Total
Full	100% (5/5)	100% (6/6)	100% (22/22)	93% (13/14)	98% (46/47)
0.1X	100% (5/5)	100% (6/6)	100% (22/22)	100% (14/14)	100% (47/47)

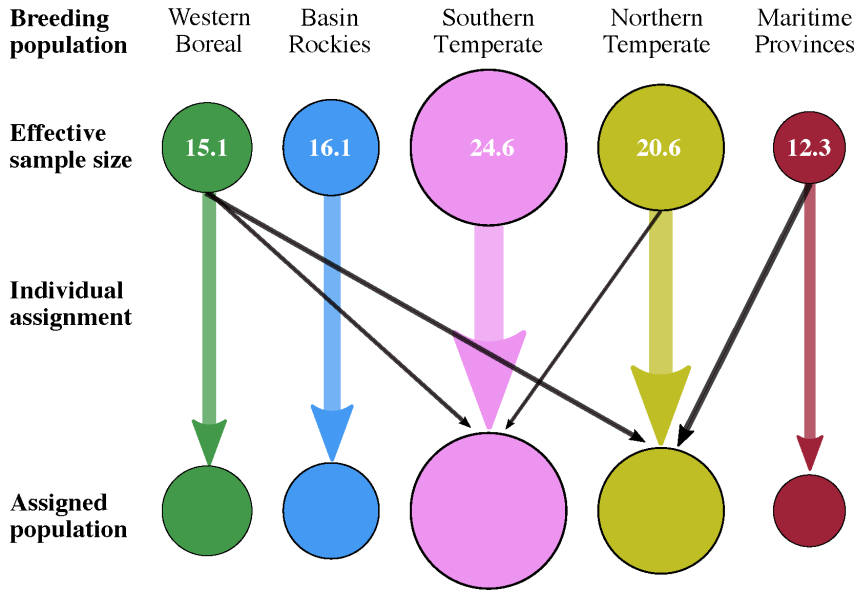
Depth	Western Boreal	Basin Rockies	Southern Temperate	Northern Temperate	Total
0.01X	100% (5/5)	100% (6/6)	95% (21/22)	100% (14/14)	98% (46/47)

Table 2. Demographic analysis of Breeding Bird Survey data for breeding populations of American Redstart.

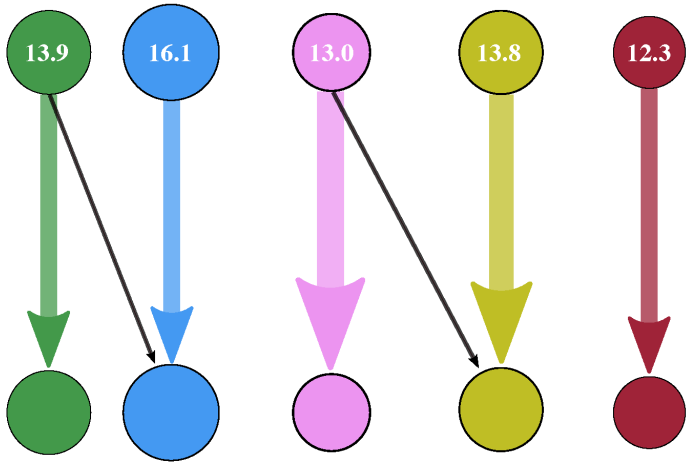
Population	No. of BBS routes	Average trend (95% CI)	Relative abundance index (95% CI)	R
Western Boreal	228	-0.90 (-1.54, -0.27)	0.66 (0.52, 0.84)	1.
Basin Rockies	84	-2.38 (-3.40, -1.30)	0.01 (0.01, 0.02)	0.
Southern Temperate	673	-0.34 (-0.75, 0.06)	0.15 (0.13, 0.17)	0.
Northern Temperate	615	0.67 (0.33, 1.01)	3.70 (3.10, 4.52)	15
Maritime Provinces	166	-1.26 (-1.73, -0.77)	1.97 (1.57, 2.49)	30



A) All samples



B) Sample size standardized



C) Effective size standardized