UNIVERSITY OF CALIFORNIA
RIVERSIDE

A Bird's Eye View of Speciation with Gene Flow: Insights from Genetic Clines Across the Yellow-Rumped Warbler Hybrid Zone

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Biology

by

Daniel Pierce

June 2024

Dissertation Committee:
Dr. Alan Brelsford, Chairperson
Dr. Polly Campbell
Dr. Kate Ostevik

The Dissertation of Daniel Pierce is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

Thank you to:

The Alberta Conservation Association, which funded fieldwork and sample collection.

The University of California, Riverside High-performance Computing Cluster, which is facilitated by grants from the NSF and the NIH.

The University of California, San Diego Institute for Genomic Medicine sequencing services for generating the sequence data used for these analyses.

Others that contributed to sample collection and offered useful insight and feedback on these projects:

Darren Irwin	David Toews
Alana Demko	Elisa Henderson
Olga Lansdorp	German Lagunas-Robles
Michael Moretti	Marie Palanchon
Allison Patterson	Sam Yeaman

DEDICATION

This work is dedicated to Gi Hyun Pierce and Randall James Pierce.

Thank you for your sacrifice on my behalf, for laying the foundation of my academic success, and sowing the seeds of my interest in science, art, and adventure.

ABSTRACT OF THE DISSERTATION

A Bird's Eye View of Speciation with Gene Flow: Insights from Genetic Clines Across the Yellow-Rumped Warbler Hybrid Zone

by

Daniel Pierce

Doctor of Philosophy, Graduate Program in Evolution, Ecology, and Organismal Biology
University of California, Riverside, June 2024
Dr. Alan Brelsford, Chairperson

When divergent populations come into contact and interbreed, incompatible genotypes can result in decreased hybrid fitness and partial reproductive isolation. By studying patterns of gene flow between populations, we can identify the genetic, phenotypic, and ecological components of partial reproductive isolation. Using whole-genome sequence data generated from 1201 yellow-rumped warblers (*Setophaga coronata coronata*, *S. c. auduboni*, and their hybrids), we measure variation of a hybrid zone's position in space and over time, identify genomic regions where gene flow is restricted, and compare the genetic basis of reproductive barriers to the genetic basis of plumage traits. We find that reproductive isolation is generated by the effects of many loci throughout the genome, with a large influence of the sex-chromosomes, that "barrier loci" cluster in some highly differentiated regions, and that gene flow and introgression are broadly asymmetric. Barrier loci show strong linkage disequilibrium, which augments the strength of the barrier to gene flow and may indicate the presence of genetic incompatibilities. Clines in plumage traits and their associated SNPs suggest strong selection against some hybrid phenotypes, although loci with the strongest associations with plumage color traits do not exhibit the strongest barrier to gene flow.

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 1

Chapter 2

Chapter 3

LIST OF TABLES

Introduction

Speciation, the process by which biological lineages diverge and become reproductively isolated, is a theme with central importance in evolutionary biology, as it is a key process in the generation and maintenance of biodiversity. Despite a long history of study, including theoretical, experimental, and observational methods, many of the details of this process remain opaque to researchers. Determining the necessary conditions for speciation with gene flow is especially challenging, since admixture of populations homogenizes genetic variation as recombination counteracts the establishment or maintenance of divergent genomes and the genetic components of reproductive isolation. Recent theoretical work exploring this problem has determined that the requirements may not be particularly stringent. Divergent selection on many genetic loci and the basic organization of genes into genomes can influence the evolutionary trajectory of interbreeding populations and tip the balance between selection and recombination in favor of speciation (Flaxman, Wacholder, Feder & Nosil 2014; Nosil, Feder, Flaxman & Gompert 2017). Connecting theory to reality is challenging however, because the empirical signal of reproductive isolation in the genomes of interbreeding populations is obscured by myriad evolutionary processes that may generate confounding patterns (Noor & Bennett 2009).

As large-scale population genomic data have become more accessible, researchers are able to refine methods for detecting the genomic signatures of speciation with gene flow. Hybrid zones, geographic regions where divergent populations interbreed, feature prominently in empirical studies of speciation with gene flow. Hybrid zones represent

intermediate stages of speciation, where the boundaries between species are “porous” due to incomplete reproductive isolation (Wu 2001). The emerging picture is that speciation is a complex process that can involve many genetic loci and involve competing and simultaneous processes affecting gene flow differentially across the genome (Dusfresnes et al. 2017; Nikolakis et al. 2022). Such empirical work has begun to bridge the gap between theory and reality. Studying hybrid zone dynamics in a broad range of taxa with differing biogeography, demography, reproductive strategies, and genome structure will provide greater insight into the processes that drive speciation with gene flow and the genomic patterns generated by the competing action of recombination and selection.

The work presented below follows this tradition by examining a secondary contact hybrid zone between the breeding ranges of two subspecies of yellow-rumped warbler, *Setophaga coronata coronata* and *S. c. auduboni*. These subspecies are phenotypically and genetically differentiated, and interbreed extensively where their populations intersect. Genetic and phenotypic distinction has been maintained over a long period of time despite ongoing hybridization (Brelsford & Irwin 2009; Mila, Smith, & Wayne 2007; Hubbard 1969). The following study aims to assemble a genome-scale view of the process of speciation with gene flow by elucidating the environmental and demographic context of the hybrid zone, the genetic architecture of reproductive barriers, and the contribution of divergent sexual signals to reproductive isolation. The accuracy of our approach is enhanced by a large collection of whole-genome sequence data from 1201 yellow-rumped warblers that span the continuum of ancestry between parent populations. We combine well-established and novel techniques for analyzing gene flow

to provide insight into the genomic signature of reproductive isolation and a basis for further studying reproductive barriers at the gene level.

This study has three primary objectives. First, we aim to determine whether the hybrid zone is stable over time and consistent across space by comparing gene flow across multiple transects sampled at two time points. The stability of a hybrid zone depends on the extrinsic and intrinsic factors that mediate the balance between selection against hybrids and gene flow via dispersal and can provide information regarding the factors maintaining or eroding species boundaries (Barton & Hewitt 1985). Next, we attempt to identify the genetic loci involved in reproductive barriers and compare genome-wide patterns in the distribution and associations among these "barrier loci" to predictions based on recent theory. Our extensive dataset enables us to accurately detect loci that experience restricted, excess, and asymmetric gene flow, and contrast the signature of reproductive barriers against the signal of neutral introgression. Last, we aim to build upon previous work toward identifying the genetic basis of divergent plumage coloration, which may act as a sexual signal and contribute to speciation (Brelsford, Toews & Irwin 2017). By comparing the genetic architecture of reproductive isolation to the genetic architectures of traits potentially involved in reproductive success, we can determine whether these traits contribute to a barrier to gene flow and move beyond inferences based solely on genetic or phenotypic differentiation, which may or may not be indicative of barrier loci.

References

Brelsford, A., Toews, D. P., & Irwin, D. E. (2017). Admixture mapping in a hybrid zone reveals loci associated with avian feather coloration. Proceedings of the Royal Society B: Biological Sciences, 284(1866), 20171106.

Dufresnes, C., Brelsford, A., Jeffries, D. L., Mazepa, G., Suchan, T., Canestrelli, D., ... & Crochet, P. A. (2021). Mass of genes rather than master genes underlie the genomic architecture of amphibian speciation. Proceedings of the National Academy of Sciences, 118(36), e2103963118.

Flaxman, S. M., Wacholder, A. C., Feder, J. L., & Nosil, P. (2014). Theoretical models of the influence of genomic architecture on the dynamics of speciation. Molecular ecology, 23(16), 4074-4088.

Hubbard, J. P. (1969). The relationships and evolution of the Dendroica coronata complex. The Auk, 86(3), 393-432.

Mila, B., Smith, T. B., & Wayne, R. K. (2007). Speciation and rapid phenotypic differentiation in the yellow-rumped warbler Dendroica coronata complex. Molecular Ecology, 16(1), 159-173.

Nikolakis, Zachary L., et al. (2022). Evidence that genomic incompatibilities and other multilocus processes impact hybrid fitness in a rattlesnake hybrid zone. Evolution 76.11, 2513-2530.

Noor, M. A., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity, 103(6), 439-444.

Nosil, P., Feder, J. L., Flaxman, S. M., & Gompert, Z. (2017). Tipping points in the dynamics of speciation. Nature ecology & evolution, 1(2), 0001.

Wu, C. I. (2001). The genic view of the process of speciation. Journal of evolutionary biology, 14(6), 851-865.

Chapter 1: The yellow-rumped warbler hybrid zone exhibits little variation across space and time

Summary

Variation in clines across a hybrid zone over time or across space can provide clues about the types of reproductive barriers that lead to divergence between populations. A cline that is stable in its position over time and consistent in its shape at different points across a hybrid zone is likely maintained by factors that are independent of the environment, such as, genetic incompatibilities that result in sterility or reduced viability. In contrast, movement of a cline over time can indicate asymmetries in dispersal or selection, or that ecological divergence is an important reproductive barrier if cline movement coincides with a changing ecotone. Variation in clines across space can also suggest a role of ecology if the hybrid zone is sampled across environmentally variable transects. This study estimates the position of four clines across the yellow-rumped warbler (*Setophaga coronata ssp.*) hybrid zone and compares these positions across two time points. We find that the hybrid zone is stable over time and shows little variation across space despite substantial change in the environment over the past two decades due to an outbreak of the mountain pine beetle (*Dendrdoctonus ponderosae*) that has affected some regions within the hybrid zone. These results suggest that maintenance of reproductive isolation between *S. c. coronata* and *S. c. auduboni* primarily depends on intrinsic reproductive barriers rather than factors that vary across space and time.

Background

When divergent and partially reproductively isolated populations interbreed, a hybrid zone may form, where individuals are genetically and phenotypically intermediate between the two parent populations. If hybrids experience reduced fitness relative to individuals of the parent populations, the location, size, and persistence of the hybrid zone depends on a balance between selection against hybrids and dispersal into the contact zone. This type of hybrid zone may be described as a "tension zone," alluding to the opposing effects of selection that reduces effective gene flow and dispersal that promotes admixture (Barton & Hewitt 1985). By characterizing the stability of a hybrid zone over time and variation in the genetic cline across multiple transects, researchers can gain insight into the evolutionary, ecological, and demographic factors that affect the distributions of species and the process of speciation. For example, asymmetries in selection and/or dispersal are implicated in the movement of several butterfly hybrid zones (reviewed in Martins, Warren, McMillan, & Barett 2023). Also, in hybrid zones mediated by ecological differences between species, changes in the location of an ecotone over time can drive rapid shifts in the location of the hybrid zone, as seen in chickadees, following changes in a climate gradient (Taylor et al. 2014) and in lizards, following alteration of habitat in part of their ranges (Leaché, Grummer, Harris & Breckheimer 2017).

Hybrid zone movement has been inferred from a variety of lines of evidence, including asymmetries in introgression across a cline (Sequeira et al. 2022) or direct comparisons of populations between time points (Wang, Delmore & Irwin 2019; Leaché,

Grummer, Harris & Breckheimer 2017). Patterns of gene flow estimated from a single time point may be consistent with more than one hypothesis, so the most compelling evidence of hybrid zone movement comes from direct measurements of a hybrid zone across time periods. For example, movement of a hybrid zone between Townsend's and Hermit warblers was inferred from a genetic "wake" of Hermit warbler mitochondrial DNA left in a region currently inhabited by Townsend's warblers (Krosby & Rowher 2010; Rowher, Birmingham, & Wood 2001). In a re-analysis of this hybrid zone, where each of the historically sampled sites were resampled for a direct comparison, evidence suggests that this hybrid zone is stable and that if movement has occurred, it occurred in the distant past following range expansion coincident with glacial retreat (Wang, Delmore & Irwin 2019).

Yellow-rumped warblers include several subspecies distributed across the North American continent. This study focuses on the myrtle (*Setophaga coronata coronata*) and Audubon's (*S. c. auduboni*) subspecies, which breed in coniferous forests of Canada and the United States of America and form a hybrid zone during their breeding season that extends from the pacific coast of British Columbia to Montana, roughly following the Rocky Mountains (Figure 1.1). This hybrid zone was initially described in detail from samples collected in 1965 (Hubbard 1969) by measuring a hybrid index of plumage traits along two transects that cross the hybrid zone. The genetic and phenotypic clines between *coronata* and *auduboni* subspecies of yellow-rumped warblers were examined 40 years later (Brelsford and Irwin 2009), by re-sampling localities Hubbard documented along two transects and adding three additional transects that cross the hybrid zone. The 2009

study found a narrow cline Brelsford and Irwin also inferred stability of this hybrid zone by comparing Hubbard's hybrid index to hybrid indices based on plumage and few molecular markers, finding that the clines reported in 1969 are approximately coincident. However, the low resolution of the plumage-based hybrid indices used by Hubbard and Brelsford and Irwin, and the small number of genetic markers used by Brelsford and Irwin limits the accuracy and precision of estimates of ancestry in admixed populations so subtle differences among transects or across time may be obscured.

A climate-change driven outbreak of the mountain pine beetle (*Dendroctonus ponderosae*), which peaked around the year 2004, has caused extensive damage to the coniferous forest within British Columbia and has since begun to spread eastward into the forests of Alberta, Canada (Safranyik & Wilson 2007). The regions most severely affected by the outbreak include a large portion of the *auduboni* breeding range near and within the hybrid zone. If the outbreak affects demography and dispersal across or into the hybrid zone, the effect is likely asymmetric and should be detectable as a change in the center of a cline at transects that cross outbreak-affected regions. For example, if *auduboni* populations respond negatively to the outbreak conditions, we might expect a decrease in local population density near the hybrid zone, resulting in lower gene flow from *auduboni* to *coronata* in the affected transect, and so a shift in the center of the genetic cline toward *auduboni* populations – in other words, a shift in the direction of greater dispersal (Barton & Hewitt 1985). Comparing current populations to populations sampled during the peak of the outbreak may provide some insight into whether

temporally variable ecological processes affect hybrid zone dynamics of yellow-rumped warblers.

The yellow-rumped warbler hybrid zone also presents an opportunity to study the influence of factors that vary across the landscape. This hybrid zone covers a great distance along the Canadian Rocky Mountains, and interbreeding occurs within valley corridors across the mountain range. These corridors vary in latitude, the extent of alpine tundra unsuitable for yellow-rumped warbler breeding, and so may vary with respect to their contributions to a geographic barrier. If spatially variable ecological factors influence reproductive isolation, we may expect that the slope of genetic clines across different transects will vary in accordance with the strength of the barrier imposed by each corridor. For example, a weaker barrier to gene flow will correspond with a change in the genetic hybrid index over a greater distance, producing a cline with a relatively shallow slope. Understanding the process of speciation requires knowledge about the extrinsic and intrinsic factors that contribute to reproductive isolation. Brelsford & Irwin (2009) ruled out assortative mating as a component of the partial reproductive isolation between *coronata* and *auduboni*, but a more detailed analysis is required to determine the roles of ecological adaptation, intrinsic genetic factors, and other behavioral aspects of reproduction.

The present study aims to determine if temporally or spatially variable aspects of the environment influence reproductive isolation between *Setophaga coronata coronata* and *S. c. auduboni*. We re-examine the yellow-rumped warbler hybrid zone in much greater detail than previous studies, using whole genome resequencing to generate data

from samples collected for the 2009 study, recent repeated sampling along the same transects with greater attention to the edges of the hybrid zone, and adding dozens of additional samples from allopatric populations. This larger data set improves the scope and resolution of sampling, increasing the number of individuals and genomic loci, to allow precise estimates of ancestry for each individual. We analyze clines in whole-genome ancestry to determine if the hybrid zone has shifted in the years since sampling was last conducted between 2005 and 2007, and in which direction it may have moved, following the mountain pine beetle outbreak. We also aim to determine whether gene flow is affected by spatial variables by comparing the estimated center of clines among several transects across the hybrid zone. This study will contribute to a much greater understanding of the components of the environment that maintain partial reproductive isolation and provide insight into the susceptibility or resilience of avian populations to indirect effects of climate change.

Methods

Sample collection

Our data consist of samples from multiple sources. Our primary samples are blood samples collected from yellow-rumped warblers in the field during the summers of 2005 – 2007, 2017, and 2019 – 2021. These samples are categorized as "historical" (2005 – 2007; n = 558) or contemporary (2017 – 2021; n = 555). To supplement our sampling of distant allopatric sites, we included published genome sequence data from New Mexico (n = 10; Szarmach, Brelsford, Witt & Toews 2021) and New York (n = 5; Baiz et al. 2021), as well as field collections from Alaska and California between 2013 and 2020

(n = 81). These allopatric samples are included in both time periods for analysis of the hybrid zone, with the exclusion of two birds sampled from California during the non-breeding season. These two birds were included for the purpose of estimating genome-wide ancestry, however.

Original wild-caught samples comprise 5 transects (Willow, Pine, Jasper, Blaeberry, and Kananaskis – listed from northern- to southernmost) that cross the hybrid zone. In addition, we acquired samples from allopatric sites far from the hybrid zone. Wild-caught birds from our contemporary data set were located by searching along public roads in their breeding habitat. Once found, birds were lured into a 6 meter-long mist net using audio playback of adult songs and calls as well as nestling calls. We banded contemporary birds with numbered aluminum leg bands. We collected blood for our genetic samples by brachial venipuncture. Last, we measured several morphometric traits (body mass, wing chord length, bill length, bill width, bill depth, tarsus length) and photographed each bird for scoring plumage traits. Each bird was released at the capture site after handling. Wild-caught historical birds were captured and processed by identical methods but a subset of these birds were given colored leg bands (detailed methods in Brelsford & Irwin 2009). Blood samples were suspended in 0.5 mL Queen's Lysis Buffer in 1.7 mL microcentrifuge tubes and stored at -20 degrees Celsius prior to extraction. Four of the five transects (Pine, Jasper, Blaeberry, and Kananaskis) were sampled sufficiently in both historical and contemporary periods, so these four transects are the focus of this study.

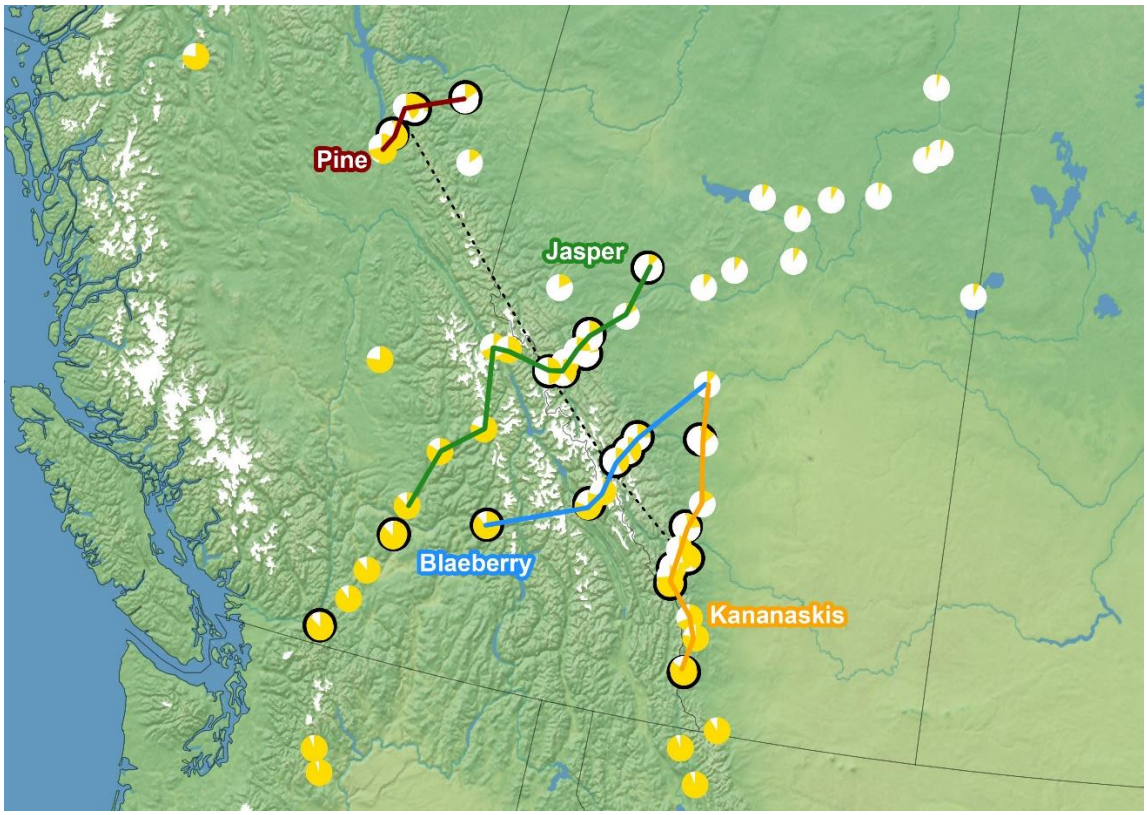

Figure 1.1: Map of sampling localities for four focal transects. Black dashed line indicates the approximate midpoint of the hybrid zone. Pie charts indicate the average proportion of genetic ancestry from *coronata* (white) and *auduboni* (yellow). Black outlines around pie charts indicate that a locality was sampled in both historical and contemporary time periods. Sites without a black outline may be historical or contemporary samples. Additional details provided in Table S1.1.

Sample processing and library preparation

We added 15 µL of 20% SDS and 13 µL of proteinase-K to each tube containing blood samples and incubated them overnight (12 to 18 hours) at 56 degrees Celsius. We manually extracted 48 samples, following Qiagen's standard protocol for the Qiagen DNEasy blood & tissue kit. For the manual extractions, we used 200 to 250 µL of the digested original sample and eluted with 90 µL of 10 mM Tris-HCL, pH 8.0. The remaining samples were extracted using the Qiacube HT, following Qiagen's recommended protocols. For the Qiacube HT extractions, we used 30 µL of each original

digested sample, diluted to a volume of 200 µL with 170 µL of Queen's Lysis Buffer, pH 7.5. We eluted the purified DNA with 90 µL of 10 mM Tris-HCL, pH 8.0 and stored at -20 degrees Celsius prior to library preparation.

We normalized the concentration of purified DNA samples to an average concentration of 1.7 ng/µL across sets of 8 samples on each plate by first measuring the concentration of each sample via Qubit, then diluting each sample with an appropriate volume of 10 mM Tris-HCL, pH 8.0. Our dilutions ensured that no diluted sample concentration exceeded a maximum of 4.2 ng/µL or fell below the minimum of 0.83 ng/µL, as specified by the Seqwell Plexwell 384 LP whole-genome library preparation kit. We followed the standard Plexwell 384 LP protocol for all library preparations. To obtain our target sequencing depth while minimizing the potential for sequencing PCR duplicates, we prepared two replicate libraries of each sample using different i7 barcodes. Specifically, each 96 well Sample Barcode plate was prepared with 48 samples, each of which was duplicated on the 96 well plate. For example, sample "AF12B01" would be added to well A1 and A7 of a sample barcode plate, sample "AF12B02" would be added to well B1 and B7 of a sample barcode plate, and so on.

Sequencing

We shipped our samples to the University of California, San Diego Institute for Genomic Medicine for 150 bp paired-end sequencing on the NovaSeq S4 platform. We sequenced 192 libraries, representing 96 individual birds (each bird replicated), on 1 lane in order to obtain an approximate sequencing depth of 3.5X per library, or 7X per individual bird.

Public database

Reads for 15 yellow-rumped warblers were acquired from the NCBI Sequence Read Archive (Baiz et al. 2021).

Data preparation

We used PEAR v0.9.10 (Zhang, Kobert, Flouri, & Stamatakis, 2014) to merge paired end reads (including the -k option to preserve alignment orientation to prevent inappropriate read dropping by the alignment tool) removed PCR duplicates with SAMtools v1.14 (Li et. al., 2009; Li, 2011), then aligned reads for each individual bird to the chromosome-level Myrtle warbler reference genome (Baiz et al. 2021) with BWA v0.7.17 (Li & Durbin, 2009). We then merged replicate sequence data for each individual and called SNPs with SAMtools v1.14. After calling variants, we used VCFtools (Danecek et al. 2011) to apply filters to exclude indels, sites with more than two alleles, and set a minimum minor allele frequency of 0.05. We computed statistics on missing data and depth on a per-site and per-individual basis to determine filtering thresholds that maintain the maximum number of informative sites. We excluded individuals with greater than 90 % missing data and excluded sites with greater than 80 % missing data or a depth less than 2 or greater than 40. Samples originating from populations in New York and New Mexico were obtained from the Sequence Read Archive and had relatively high proportions of missing data relative to our original libraries. These samples are the only representatives of genetic variation in the southwest and northeastern United States, so the permissive filtering allows us to retain some information about allele frequencies in

the parental populations. Given the large number of SNPs in the resulting dataset, even if only 20% of variants are present for a sample, this still includes millions of SNPs.

Population structure and individual ancestry

In order to address the non-independence of linked loci in our estimation of individual ancestry, we pruned our full data of SNPs with strong allelic correlations. We computed pairwise r^2 for a randomly selected subset of 10 % of the full data with PLINK v1.90b6.25 (Purcell et al. 2007) and plotted r^2 against the distance between loci in a pair. We pruned one of each locus in a pair with r^2 greater than 0.1 (the approximate level of background LD) for distances of less than 50 SNPs. We then used Admixture v1.3.0 (Alexander, Novembre, & Lange, 2009) to infer ancestry of each sample given 2 putative populations (K = 2).

Spatial analysis

In order to perform our analyses of the hybrid zone clines, we averaged sample coordinates and ancestry proportions (via Admixture) for each locality and plotted these ancestry and coordinate averages in QGIS v3.26 (QGIS.org 2024). We then identified localities for each transect where the average genome-wide ancestry is greater than and less than 0.5, then interpolated between these localities to find the theoretical midpoint where average ancestry is exactly 0.5. With the midpoints of each transect defined, we fit a line through the midpoints of the Kananaskis, Blaeberry, Jasper, and Pine transects to represent a continuous hybrid zone across the sampling range. We obtained the distance between each sampling locality and the nearest point along the line through the hybrid zone midpoints using the “Distance to Nearest Hub” tool in QGIS. We did not obtain

enough samples from the Willow transect for our contemporary data set to warrant inclusion in this analysis.

Cline analysis

We fit clines to genome-wide ancestry averaged among each locality using HZAR v0.2-5 (Derryberry, Derryberry, Maley & Brumfield 2014). HZAR estimates several parameters that describe the position of the cline along a transect where a trait value or allele frequency is expected to be exactly 0.5 with respect to parental populations, "center," and the "width" of the cline defined as 1/ maximum slope (Derryberry, Derryberry, Maley & Brumfield 2014). We fit clines to each transect at each period individually and treated ancestry as a quantitative trait. Initial testing showed that fitting cline models with no tail parameters estimated performed poorly, so we retained fitted clines only for models estimating mirrored tails or both tails estimated separately. For comparison of clines across time periods, we chose the model for each transect in each period with the lowest corrected AIC score. To determine if the hybrid zone center has shifted since our historical sampling, we generated a separate fit, using the same model, that constrained the contemporary cline to the center point estimate from the historical cline fit. We compared corrected AIC scores for this constrained contemporary cline and the unconstrained contemporary cline and determined that a difference of 2 or more indicated a significant difference in cline center. If we determined that a cline was significantly different between the historical and contemporary sampling of a transect, we also fit a cline to a reduced contemporary data set for that transect, where the birds included in the analysis were pruned to resemble the total number and spatial distribution

of birds in the historical sampling of the transect. This follow-up analysis was done to rule out the effects of the differences in sampling effort between time periods, since contemporary sampling included many more samples from the edges of the hybrid zone, and historical sampling focused primarily on sites within the hybrid zone.

Results

Whole genome libraries

After filtering our raw sequence data to remove individuals with a high proportion (> 90%) of missing data and birds that appeared to have high relatedness (relatedness > 0.3, representing likely siblings) to another bird in our dataset, we retained 1201 of our 1214 sequenced samples. Alignment rate for the reads we obtained is, on average, 84 %. Per-individual average sequence depth is approximately 6X.

Genome-wide ancestry clines

To test for a change in the location of the hybrid zone center, we constrained the contemporary cline model to the center estimate from the corresponding historical cline. The fit of the constrained cline was similar for all but the Jasper transect (Table 1.1, Figure 1.2). The difference in cline center between the historical and contemporary Jasper transect is 10.3 km, where the fit of the constrained model is substantially worse (ΔAICc = 92). To determine whether differences in sampling can account for the change in cline center, we pruned the contemporary Jasper transect so that the total number and distribution of individual birds used for the cline analysis resembled that of the historical transect and performed cline fitting on this pruned transect (Figure 1.3). The pruned contemporary transect through Jasper had a width of 199 km (95% confidence interval:

150 km, 235 km) and a center at -5.8 km (-15, -1.8). The 95% confidence intervals for the cline center obtained using the pruned and full data overlap (full contemporary transect: -12.2, -11.6; pruned contemporary transect: -14.7, -1.8) so it is unlikely that sampling differences account for the unusual characteristics of the cline through Jasper, when compared to the other transects.

Transect	**Period**	**Center**	**Center CI**	**Width**	**Width CI**	**Model**
Kananaskis	Historical	-0.58	(-6.28, -0.58)	135	(135, 157)	Mirror
Kananaskis	Contemporary	-5	(-6.9, -3.9)	146	(146, 151)	Both
Blaeberry	Historical	3.6	(-1.4, 9.1)	197	(178, 234)	Mirror
Blaeberry	Contemporary	-3.1	(-8.8, 3)	154	(117, 191)	Mirror
Jasper	Historical	-1.3	(-6.9, 7)	220	(189, 269)	Mirror
Jasper	Contemporary	-11.6	(-12.2, -11.6)	236	(233, 236)	Both
Pine	Historical	-2.1	(-6.5, -1.3)	135	(129, 154)	Both
Pine	Contemporary	-5.8	(-15.1, 3.8)	144	(99, 176)	Both

Table 1.1: Summary of cline parameters for each transect and time period. 95 % confidence intervals for cline parameters given in parentheses. Model refers to whether HZAR was instructed to estimate cline tail parameters for the western and eastern tails individually (both) or as the same value but opposite direction (mirror).

Cline width showed no significant change over time for any transect. The cline across Jasper exhibits an unusually wide cline, where the historical estimate is 220 km (189, 269) and the contemporary estimate is 236 km (233, 236), nearly 80 km wider than the average of all other transects (Table 1.1, Figure 1.2). The historical Blaeberry cline is also relatively wide at 197 km (178, 243), while the contemporary cline is much narrower at 154 km (117, 191). The confidence intervals for both the historical and contemporary clines fit to the Blaeberry transect are large, so we tested for a change in cline width over time by constraining the width of the contemporary cline to the width estimate of the historical cline and found no support for a change in width (ΔAICc = 1).

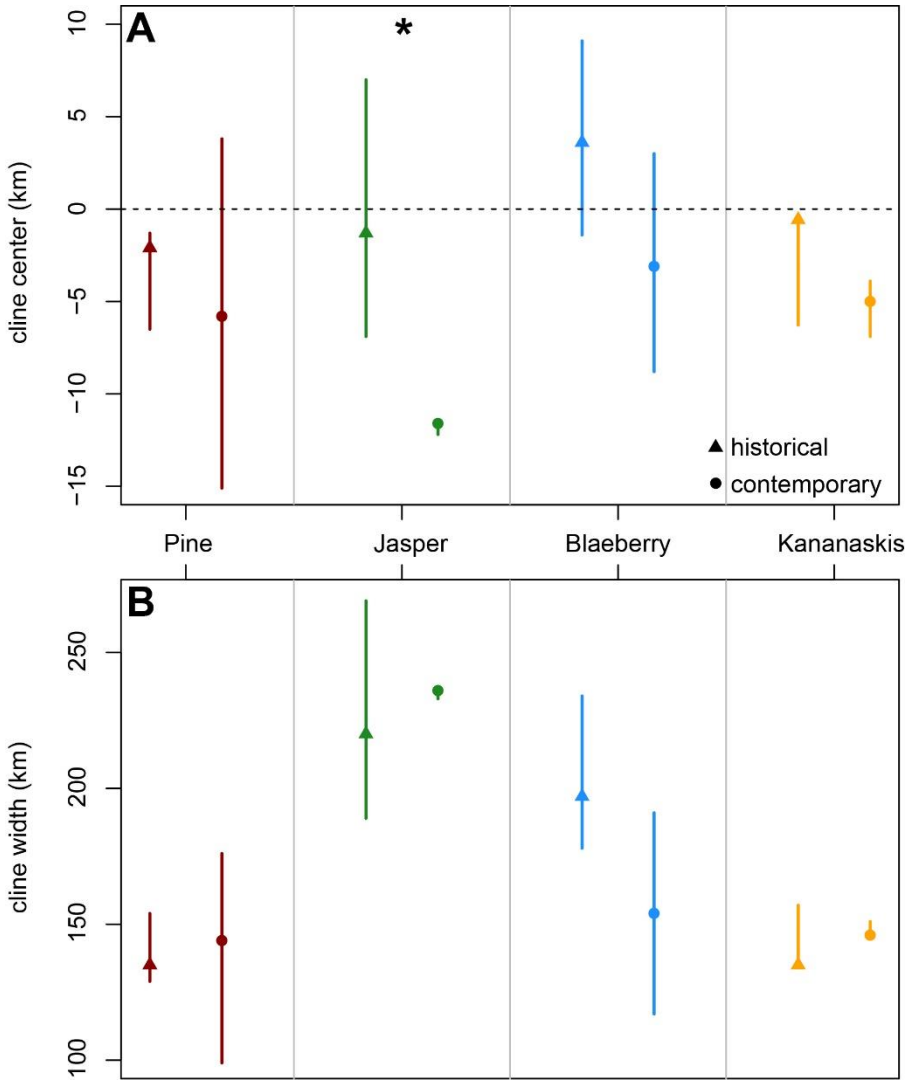

Figure 1.2: Summary of genome-wide ancestry cline parameters by transect and time period. Transects are grouped by color. The point estimate of a cline is represented by points (triangles for historical and circles for contemporary) and 95% confidence intervals are shown as vertical lines through a point. Panel A shows cline center estimates and cline width estimates are shown in panel B. Asterisk indicates a significant difference between a parameter estimate between periods.

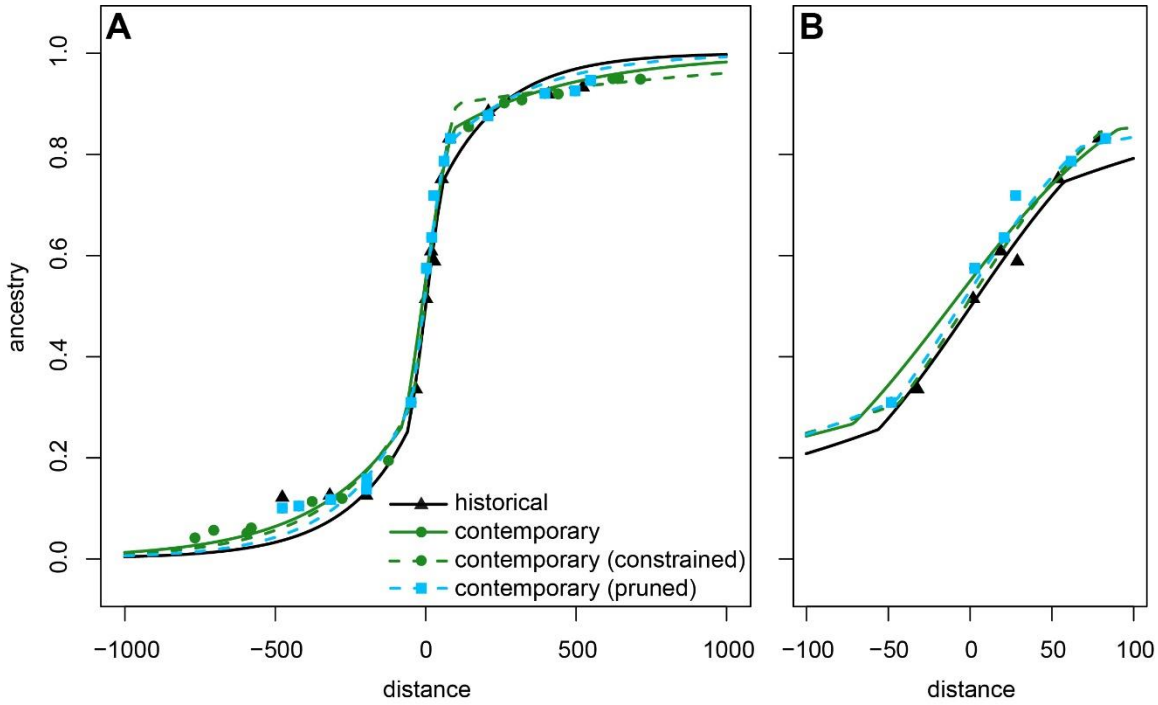

Figure 1.3: Clines fit to the historical sampling (black solid line) contemporary sampling (solid green line), contemporary cline constrained to the historical center (dashed green line) and pruned contemporary sampling (dashed blue line) of the Jasper transect. Points indicate the localities sampled along the transect and point colors correspond with the lines of the same color. Panel A shows the central 2000 km portion of the transect and panel B shows the central 200 km portion.

Discussion

We analyzed clines in genome-wide admixture to determine whether a hybrid zone in yellow-rumped warblers has shifted between 2007 and 2021, and whether there are differences in cline shape and position between four transects that cover most of the geographic extent of the hybrid zone. Across all transects, the point estimate of cline center decreases slightly, with an average difference of -6.3 km, or an average shift of 6.3 km westward. However, the change in cline center is significant only in the Jasper transect, where the cline has moved westward by 10 km. We do not detect a change in cline width over time for any of our transects.

The change in cline center in the Jasper transect is relatively small (0.6 km per year) compared to the estimated width of the cline estimated at the Jasper transect (210 km) and relative to the potential for dispersal. Brelsford & Irwin (2009) estimated that the per-generation dispersal distance for yellow-rumped warblers is 20 (13 – 25) km. Other studies that directly compare the location of hybrid zones across two time periods demonstrate that hybrid zones of similarly mobile animals can move at a slightly higher rate, approximately 1 km per year (Aguillon 2022; Taylor et al. 2014; Ryan et al. 2018. Given that the other three transects are consistent in their positions over time, it is possible that the shift we observe in Jasper is due to either sampling error or subtle asymmetries in dispersal or selection. As an additional check for subtle changes in the hybrid zone, we compared the average of local ancestry at resampled sites from within or near the hybrid zone and found no consistent pattern of change (Figure 1.4).

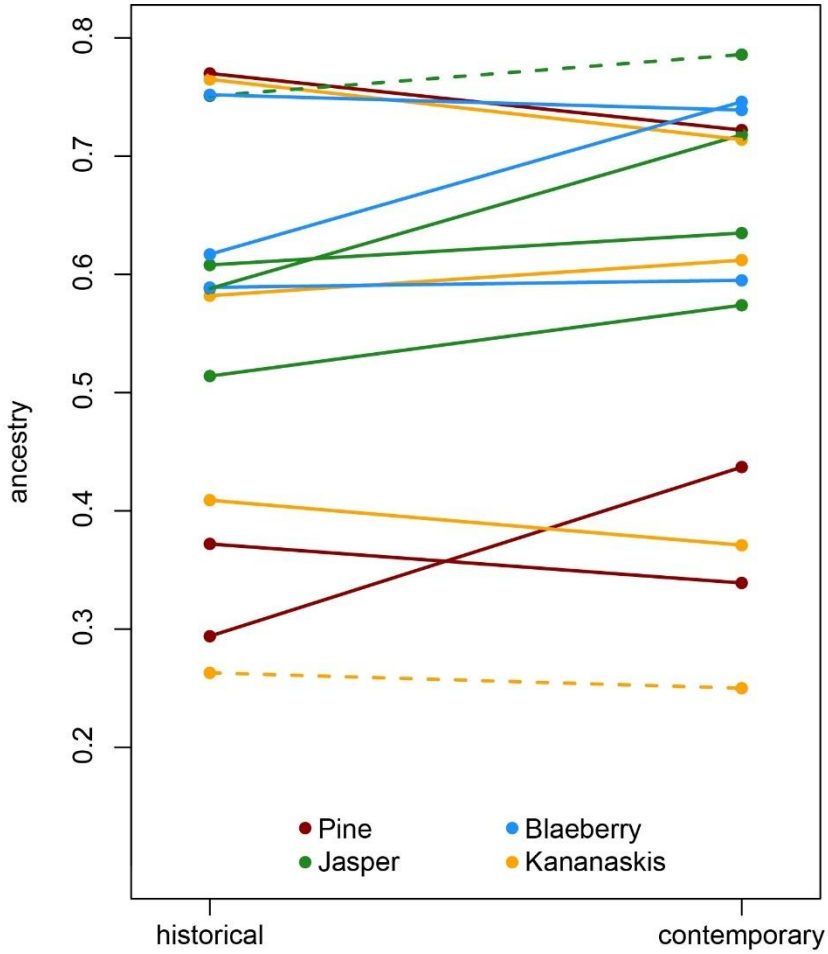

Figure 1.4: Average ancestry at sites sampled in both historical and contemporary time periods. Dashed lines indicate that the contemporary sampling locality is more than 5 km from the historical locality.

Warmer winters have contributed to lower yearly mortality and range expansion of the mountain pine beetle, which has led to the destruction of vast stands of trees within habitat that would otherwise support the *audubonii* population during their breeding season. The mountain pine beetle has spread into northwestern regions of central Alberta (Robertson et al. 2009), but the damage has been less severe there than in British Columbia. Whether a pine beetle outbreak is beneficial, neutral, or deleterious for wildlife depends on the taxa considered; for birds in particular, nesting strategy appears

to be an important factor for making predictions (Mosher et al. 2019; Janousek et al. 2019; Saab et al. 2013). Although yellow-rumped warblers are predicted to exhibit an increase in habitat use following an outbreak (Mosher et al. 2019; Janousek et al. 2019), the *coronata* and *auduboni* subspecies are not treated as unique populations and measures of occupancy do not necessarily reflect patterns of dispersal and selection, which are important aspects of hybrid zone dynamics. If substantial and asymmetric changes in the breeding habitat, as a result of the mountain pine beetle outbreak, influence the current position of the cline center, we would expect to find a shift, over time, in the transects that cross regions where the effects of the outbreak have been most severe – namely, across the Pine and Jasper transects. While the Jasper transect does show a measurable shift in cline center, we find no significant change in the Pine transect. The Kananaskis cline may be expected to show the least change over time, since this transect crosses a region along the eastern edge of the Rocky Mountains, which has sustained much less severe damage from the outbreak (Safranyik & Wilson 2007). However, the center and width of the Kananaskis cline is similar to Pine and Blaeberry, so it is unlikely that the effects of the pine beetle outbreak affect the extent and symmetry of gene flow.

Instead, the slight and statistically insignificant, albeit consistent westward shift in 3 of our transects, and the larger and better supported westward shift in Jasper implicates intrinsic factors that generate an overall asymmetry in effective gene flow across the hybrid zone. This asymmetry can be explained by greater long-distance dispersal of *coronata* or higher population density on the *coronata* side of the hybrid zone. Migratory distances of *coronata* are greater than that of *auduboni*, (Toews et. al. 2016) and their

migratory route passes over the breeding range of *auduboni*. It may be possible that some *coronata* end their migration short of their typical breeding range, and instead occupy territory in the primary breeding range of *auduboni* and *auduboni*-like hybrids, which may generate asymmetric long-distance dispersal and subtle changes in the center of the hybrid zone over time. Alternatively, asymmetric fitness effects of introgressed alleles at loci involved in reproductive barriers will result in delayed or restricted effective migration at those loci. If many loci contribute to reproductive barriers, or if there is strong linkage of these barrier loci with large regions of the genome, asymmetric selection can bias effective gene flow at a large enough proportion of the genome that it would produce a signal visible from clines in whole-genome ancestry. Determining whether asymmetry in selection contributes to the small change we observe over time requires analysis of gene flow at individual loci, and identification of which regions of the genome are involved in reproductive barriers.

HZAR also estimates cline parameters that describe the transition point between the central slope and the beginning of the tail of a stepped cline (δ), as well as the slope of the tail (τ), which can provide information about the symmetry of introgression (Derryberry, Derryberry, Maley & Brumfield 2014; Gay, Crochet, Bell & Lenormand 2008). For example, if the transition point is more distant from the center on one side of the hybrid zone, or if the slope on one side is shallower, this could suggest greater introgression on that side. For each transect, we fit a cline model that assumes symmetric tails ("mirror") and a model that fits both tails separately ("both"). The best fitting model in the historical cline for a transect was not always the best fitting model for the

contemporary transect, however the “mirror” model was preferred in 3 of 4 historical transects, while the “both” model was preferred in 3 of 4 contemporary transects (Table 1.1). We do not consider the differences in tail parameter estimates informative for this study. First, thorough sampling of the cline tails was not done in the historical sampling, so we are unable to accurately estimate the shape of the cline tails for historical transects. The lack of resolution in the cline tails for the historical transects can result in a better fit of the “mirror” model, which estimates fewer parameters. Second, variation in the cline tails between transects and on either side of a transect can be influenced by differences in topography and habitat continuity.

Concluding remarks

The stability of the yellow-rumped warbler hybrid zone over time, and its consistency across space, implicates environment-independent factors, such as between-locus genetic incompatibilities, as the primary mechanisms maintaining reproductive isolation. Genetic incompatibilities between different loci are thought to play a substantial role in the evolution of reproductive isolation and the maintenance of divergence between populations (Coyne & Orr 2004; Dobzhansky 1937; Muller 1942). In the absence of assortative mating (Brelsford & Irwin 2009), *coronata* and *auduboni* maintain genetic and phenotypic differentiation. Given this context, our results from this study suggest that intrinsic post-zygotic isolating mechanisms are important in the early stages of speciation with gene flow.

References

Aguillon, S. M., & Rohwer, V. G. (2022). Revisiting a classic hybrid zone: movement of the northern flicker hybrid zone in contemporary times. *Evolution*, *76*(5), 1082-1090.

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome research, 19(9), 1655-1664.

Baiz, M. D., Wood, A. W., Brelsford, A., Lovette, I. J., & Toews, D. P. (2021). Pigmentation genes show evidence of repeated divergence and multiple bouts of introgression in Setophaga warblers. Current Biology, 31(3), 643-649.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. Annual review of Ecology and Systematics, 16(1), 113-148.

Brelsford, A., & Irwin, D. E. (2009). Incipient speciation despite little assortative mating: the yellow-rumped warbler hybrid zone. *Evolution*, *63*(12), 3050-3060.

Coyne, J. A., & H Allen Orr. (2004). Speciation. Sinauer Associates, Oxford University Press.

Cullingham, C. I., James, P. M., Cooke, J. E., & Coltman, D. W. (2012). Characterizing the physical and genetic structure of the lodgepole pine× jack pine hybrid zone: mosaic structure and differential introgression. Evolutionary applications, 5(8), 879-891.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156-2158.

Derryberry, E. P., Derryberry, G. E., Maley, J. M., & Brumfield, R. T. (2014). HZAR: hybrid zone analysis using an R software package. Molecular ecology resources, 14(3), 652-663.

Dobzhansky, T. (1937). Genetics and the Origin of Species. Columbia university press.

Gay, L., Crochet, P. A., Bell, D. A., & Lenormand, T. (2008). Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. *Evolution*, *62*(11), 2789-2806.

Hubbard, J. P. (1969). The relationships and evolution of the Dendroica coronata complex. The Auk, 86(3), 393-432.

Janousek, W. M., Hicke, J. A., Meddens, A. J., & Dreitz, V. J. (2019). The effects of mountain pine beetle outbreaks on avian communities in lodgepole pine forests across the greater Rocky Mountain region. Forest Ecology and Management, 444, 374-381.

Krosby, M., & Rohwer, S. (2010). Ongoing movement of the hermit warbler X Townsend's warbler hybrid zone. PloS one, 5(11), e14164.

Leaché, A. D., Grummer, J. A., Harris, R. B., & Breckheimer, I. K. (2017). Evidence for concerted movement of nuclear and mitochondrial clines in a lizard hybrid zone. Molecular Ecology, 26(8), 2306-2316.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics, 25(14), 1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAMtools. bioinformatics, 25(16), 2078-2079.

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics, 27(21), 2987-2993.

Martins, A. R. P., Warren, N. B., McMillan, W. O., & Barrett, R. D. (2023). Spatiotemporal dynamics in butterfly hybrid zones. Insect Science.

Mosher, B. A., Saab, V. A., Lerch, M. D., Ellis, M. M., & Rotella, J. J. (2019). Forest birds exhibit variable changes in occurrence during a mountain pine beetle epidemic. Ecosphere, 10(12), e02935.

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. In *Biol. Symp.* (Vol. 6, p. 71).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559– 575.

QGIS.org (2024). QGIS Geographic Information System. QGIS Association. http://www.qgis.org.

Robertson, C., Nelson, T. A., Jelinski, D. E., Wulder, M. A., & Boots, B. (2009). Spatial–temporal analysis of species range expansion: the case of the mountain pine beetle, Dendroctonus ponderosae. *Journal of biogeography*, *36*(8), 1446-1458.

Rohwer, S., Bermingham, E., & Wood, C. (2001). Plumage and mitochondrial DNA haplotype variation across a moving hybrid zone. Evolution, 55(2), 405-422.

Ryan, S. F., Deines, J. M., Scriber, J. M., Pfrender, M. E., Jones, S. E., Emrich, S. J., & Hellmann, J. J. (2018). Climate-mediated hybrid zone movement revealed with genomics, museum collection, and simulation modeling. Proceedings of the National Academy of Sciences, 115(10), E2284-E2291.

Saab, V. A., Latif, Q. S., Rowland, M. M., Johnson, T. N., Chalfoun, A. D., Buskirk, S. W., ... & Dresser, M. A. (2014). Ecological consequences of mountain pine beetle outbreaks for wildlife in western North American forests. Forest Science, 60(3), 539-559.

Safranyik, L., & Wilson, B. (2007). The mountain pine beetle: a synthesis of biology, management and impacts on lodgepole pine. Canadian Forest Service.

Seneviratne, S. S., Toews, D. P., Brelsford, A., & Irwin, D. E. (2012). Concordance of genetic and phenotypic characters across a sapsucker hybrid zone. Journal of Avian Biology, 43(2), 119-130.

Sequeira, F., Arntzen, J. W., van Gulik, D., Hajema, S., Diaz, R. L., Wagt, M., & van Riemsdijk, I. (2022). Genetic traces of hybrid zone movement across a fragmented habitat. Journal of evolutionary biology, 35(3), 400-412.

Swenson, N. G., & Howard, D. J. (2005). Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. The American Naturalist, 166(5), 581-591.

Szarmach, S. J., Brelsford, A., Witt, C. C., & Toews, D. P. (2021). Comparing divergence landscapes from reduced-representation and whole genome resequencing in the yellow-rumped warbler (*Setophaga coronata*) species complex. Molecular ecology, 30(23), 5994-6005.

Taylor, S. A., White, T. A., Hochachka, W. M., Ferretti, V., Curry, R. L., & Lovette, I. (2014). Climate-mediated movement of an avian hybrid zone. Current Biology, 24(6), 671-676.

Toews, D. P., Brelsford, A., Grossen, C., Milá, B., & Irwin, D. E. (2016). Genomic variation across the Yellow-rumped Warbler species complex. *The Auk: Ornithological Advances*, *133*(4), 698-717.

Wang, S., Rohwer, S., Delmore, K., & Irwin, D. E. (2019). Cross-decades stability of an avian hybrid zone. Journal of Evolutionary Biology, 32(11), 1242-1251.

Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: a fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics, 30(5), 614-620.

Locality	Distance	n	Hybrid Index	Period	Transect
Sugarloaf	-1877	3	0.000	H	all
Sugarloaf	-1877	3	0.000	C	all
Chuska Mtns	-1735	9	0.000	H	all
Chuska Mtns	-1735	9	0.000	C	all
Alta Sierra	-1727	12	0.001	H	all
Alta Sierra	-1727	12	0.001	C	all
Parker Meadow	-1702	19	0.002	H	all
Parker Meadow	-1702	19	0.002	C	all
Sherman Pass	-1697	5	0.004	H	all
Sherman Pass	-1697	5	0.004	C	all
Mt Adams	-766	9	0.041	C	P, J, B
Spring Rimrock	-704	10	0.056	C	P, J, B
Mill Creek	-593	11	0.051	C	P, J, B
White Pine	-579	11	0.061	C	P, J, B
Skalkaho	-552	7	0.015	C	K
Beaver Creek	-481	9	0.030	C	K
Hope	-476	10	0.100	C	J, B, P
Hope	-476	12	0.121	H	J, B, P
Rattlesnake	-466	3	0.036	C	K
Britton Creek	-421	10	0.104	C	J, B
Smith	-402	10	0.041	C	K
Comstock	-377	10	0.113	C	J, B
Lac LeJeune	-318	23	0.125	H	J, B, P
Lac LeJeune	-315	9	0.117	C	J, B, P
Abbott Bay	-310	9	0.055	C	K
Oconnor Lake	-277	10	0.119	C	J
Hay Creek	-261	5	0.064	C	K
Waterton	-243	7	0.087	C	K
Yard Creek	-197	10	0.125	H	J, B
Yard Creek	-196	11	0.137	C	J, B
Clearwater	-195	8	0.158	C	J
Crowsnest	-155	11	0.149	H	K
Blue River	-123	10	0.194	C	J
Livingstone	-114	10	0.210	C	K
Etherington Creek	-86	9	0.310	C	K
Golden	-59	17	0.194	H	B
Golden	-54	11	0.212	C	B
Whiskers Point	-49	13	0.294	H	P
Red Pass Junction	-48	11	0.309	C	J

Locality	Distance	n	Hybrid Index	Period	Transect
Whiskers Point	-47	10	0.437	C	P
Peter Lougheed	-42	20	0.250	C	K
Peter Lougheed	-33	6	0.263	H	K
Mt Robson	-32	14	0.335	H	J
Blaeberry	-30	15	0.309	H	B
Kennedy	-24	11	0.339	C	P
Kennedy	-23	13	0.372	H	P
Evan Thomas	-20	19	0.409	H	K
Evan Thomas	-20	4	0.371	C	K
Sibbald West	-3	24	0.462	H	K
Rafter 6	1	7	0.523	H	K
Yellowhead Pass	2	9	0.514	H	J
Yellowhead Pass	3	9	0.574	C	J
Yamnuska	4	17	0.555	H	K
Sibbald East	9	18	0.582	H	K
Sibbald Creek	11	19	0.612	C	K
Saskatchewan RX	12	23	0.589	H	B
Saskatchewan RX	13	20	0.595	C	B
Pine West	14	14	0.571	H	P
Pine East	18	6	0.770	H	P
Pyramid Lake	19	9	0.608	H	J
Pyramid Lake	21	13	0.635	C	J
Pine Pass	22	18	0.722	C	P
Snaring	28	9	0.718	C	J
Palisades	29	15	0.588	H	J
Ghost	32	9	0.714	C	K
Waiparous	32	14	0.765	H	K
Cline River	36	10	0.746	C	B
Cline River	36	22	0.617	H	B
Pocahontas	54	8	0.751	H	J
Abraham Lake	58	13	0.752	H	B
Abraham Lake	58	11	0.739	C	B
Tumbler Ridge	58	11	0.851	H	P
Miette	62	9	0.786	C	J
Sundre	70	6	0.836	C	K
Hinton	79	11	0.831	H	J
Hinton	83	11	0.831	C	J
Moberly Lake	95	12	0.846	H	P
Moberly Lake	99	12	0.821	C	P
Rocky Mountain House	127	8	0.846	C	B, K

Locality	Distance	n	Hybrid Index	Period	Transect
Rocky Mountain House	129	11	0.867	H	B, K
Sundance	143	8	0.854	C	J
Lodgepole	186	11	0.892	C	B, K
Fox Creek	208	12	0.876	C	J
Fox Creek	209	12	0.884	H	J
McLeod	262	10	0.901	C	all
Grizzly Trail	320	10	0.907	C	all
Chisholm	397	10	0.920	C	all
Slave Lake	409	3	0.918	H	all
Fawcett Lake	441	10	0.919	C	all
Pelican Hills	497	11	0.925	C	all
Cold Lake	523	5	0.932	H	all
House River	549	9	0.946	C	all
South McMurray	623	11	0.949	C	all
Maqua Lake	642	9	0.951	C	all
Fort McKay	715	2	0.948	C	all
Anchorage	1182	26	0.992	H	all
Anchorage	1182	26	0.992	C	all
New York	3106	5	1.000	H	all
New York	3106	5	1.000	C	all

Table S1.1: Localities sampled for cline analysis comparing transects and sampling periods. Distance from hybrid zone midpoint, number of individuals representing the locality, genome-wide ancestry-based hybrid index averaged among individuals in each locality, the period of sampling applied each locality is attributed to, and the transect(s) that include each locality. Some localities in allopatric populations are sampled in only one period but included in both as they represent genetically "pure" populations. "Historical" sampling is denoted with "H" and "contemporary" sampling with "C." Transects are coded as follows: Pine – P, Jasper – J, Blaeberry – B, Kananaskis – K. Sites included in all transects represent allopatric populations and are denoted with "all" in the transect column.

Chapter 2: The genetic architecture of reproductive barriers in yellow-rumped warblers

Summary

Accurate identification of the genetic loci involved in reproductive barriers is an essential step toward characterizing the genetic basis of reproductive isolation. Our identification of "barrier loci" is supported by two analyses: cline analysis, a well-documented approach that can identify regions of the genome that contribute to selection against hybrids, and a novel approach to estimating the extent of introgression beyond the hybrid zone, which is more sensitive to the effects of individual loci and can be extended to genome regions where differentiation is low. By combining these techniques, we determine that reproductive isolation in yellow-rumped warblers is generated by the effects of many loci throughout the genome, that these barrier loci are overrepresented on the sex chromosomes, and that the barrier to gene flow is variable in its symmetry. Additionally, we show that linkage disequilibria among barrier loci are significantly stronger than linkage disequilibria among non-barrier loci, consistent with theoretical models of speciation that predict that associations among reproductive barrier loci are needed to counteract the homogenizing effects of gene flow and recombination.

Background

A recurring theme in recently developed and established theoretical models of speciation with gene flow is that statistical associations among divergently selected loci may be the most basic requirement for the establishment and maintenance of reproductive barriers (Flaxman, Wacholder, Feder, & Nosil 2014; Flaxman, Feder, & Nosil 2013; Barton & De Cara 2009; Felsenstein 1981). Selection that maintains and, over time, builds up linkage disequilibria provides a means to overcome recombination that would otherwise erode genomic divergence and reproductive barriers between interbreeding populations. Physical linkage of divergently selected genes can also facilitate speciation by protecting multilocus genotypes from recombination (Feder, Geiji, Yeaman, & Nosil 2012; Via 2012; Yeaman & Whitlock 2011; Via & West 2008) and structural features such as inversions (Feder, Nosil, & Flaxman 2014; reviewed in Huang & Rieseberg 2020) and centromeres (Carneiro, Ferrand, & Nachman 2009) can have a similar effect. Although many models that detail the conditions for speciation with gene flow focus on *in situ* speciation with gene flow, the interaction of coupled complexes of reproductive barriers is also applicable to speciation with initial divergence in allopatry followed by secondary contact. Secondary contact of divergent populations can generate initial linkage disequilibrium in the contact zone which can improve the efficiency of selection against recombinant genotypes and the maintenance of reproductive barriers, especially when the number of loci involved in barriers is large (Barton 1983; Felsenstein 1981).

Theory regarding the patterns of genomic variation that arise during the process of speciation with gene flow emphasizes the gene as the unit of reproductive isolation,

rather than whole genomes (Wu 2001). This "genic view" of speciation allows for semi-permeable species boundaries, in which loci involved in differential adaptation or reproductive isolation experience a barrier to gene flow between divergent populations, while the introgression of universally beneficial alleles or neutral loci is limited only by their linkage to selected loci (Barton 1979). The genic view of speciation is consistent with empirical observations that many pairs of taxa maintain phenotypic and genetic differentiation despite ongoing gene flow, and that genetic differentiation is often heterogeneous throughout the genome (e.g. Andrew & Rieseberg 2013; Jones et. al. 2012; Turner, Hahn, & Nuzhdin 2005). The interpretation of "islands" of differentiation as loci involved in reproductive barriers has received a great deal of scrutiny, with critics noting that the pattern can be generated by processes other than speciation (Burri et al. 2015; Cruickshank & Hahn 2014; Turner & Hahn 2010; Noor & Bennett 2009). For example, in regions where recombination is suppressed, background selection can reduce genetic diversity within populations, which can elevate differentiation even in the absence of selection against gene flow, and lead researchers to false conclusions about the involvement of these regions in reproductive isolation. Additionally, when selection on an individual barrier locus is weak, the locus may not exhibit strong differentiation between parental populations, so many regions that may be involved in reproductive barriers may evade recognition (Feder, Geiji, Yeaman, & Nosil 2012). Because reproductive barriers necessarily restrict gene flow in the corresponding genomic regions, rather than simply identifying loci that exhibit differences between populations, researchers require the means to detect regions that experience differential introgression,

specifically due to selection against recombinant genotypes. A genomic map of differential introgression between divergent populations can be used to identify the genes likely involved in reproductive barriers (Payseur 2010), and to characterize the genomic architecture of speciation.

To this end, hybrid zones are often studied as models of speciation with gene flow as they represent some of the earliest examples of semi-permeable species boundaries (Harrison & Larson 2014). Within a hybrid zone, the genomes of admixed individuals are mosaics of ancestry from the parent populations and the relative proportion of ancestry varies as backcrossing and recombination breaks down parental haplotypes. A spatially continuous hybrid zone produces a cline in genome-wide ancestry and, at individual loci, a cline in allele frequency. The width of a cline is influenced by the strength of selection against hybrid genotypes, which is balanced by dispersal of individuals into the hybrid zone, where a narrow cline corresponds with stronger selection against hybrids (Barton & Hewitt 1985; Barton & Gale 1993). The center of a cline can also be estimated, which establishes where, in space, the maximum change in allele frequency occurs. Variability in the center among clines can be used to make inferences about asymmetries in the relative strength of selection versus dispersal and spatially dependent or environmental factors maintaining a hybrid zone (Barton & Hewitt 1985).

Despite the advantages of cline analysis over genome scans for divergence or differentiation, there are limitations and drawbacks. In a simulation of hybrid zones under varying migration rate, deme size, and hybrid zone age, Jofre and Rosenthal (2021) determined that cline analysis is sensitive to the effects of drift, which increase with small

populations and low migration rates, and suggests that the demography and history of a hybrid zone be carefully considered. Additionally, identifying barrier loci from clines can be imprecise because the effect of selection on an individual locus is obscured by the effects of selection on other loci in linkage disequilibrium. In highly admixed populations, selection acts simultaneously on all loci that contribute to reproductive barriers, so the effective selection against hybrids is strong. This genome-wide selection produces a steep sigmoid curve in the center of the cline, flanked by shallower tails that represent the exponential decay of heterospecific alleles introgressing beyond the hybrid zone (Szymura & Barton 1991). To disentangle the effects of direct and indirect selection and make identification of barrier loci more precise, we can estimate selection on a locus from the tails of introgression, where many generations of backcrossing and recombination have fragmented haplotype blocks and isolated heterospecific alleles in the opposing genetic background. This approach remains untested, perhaps due to the requirement that the edges of a hybrid zone, where introgressing alleles are present only at very low frequency, are thoroughly sampled.

The yellow-rumped warbler complex is comprised of several phenotypically and genetically differentiated subspecies of migratory songbirds and includes a hybrid zone between *Setophaga coronata coronata*, and *S. c. auduboni* (hereafter *coronata* and *auduboni*). This hybrid zone likely represents a secondary contact zone formed following the last glacial maximum (Toews et al. 2016; Brelsford & Irwin 2009; Mila, Smith, & Wayne 2007; Hubbard 1969). The hybrid zone extends along the Rocky Mountains from northern British Columbia to southern Alberta, Canada, and delineates their respective

breeding ranges, where parent populations and hybrids are abundant. The hybrid zone is stable and features a steep cline maintained by selection against hybrids (Chapter 1; Brelsford & Irwin 2009; Hubbard 1969). Genetic differentiation between *coronata* and *auduboni* is very high in isolated regions of their genomes, contrasted by a background of very low differentiation across most of the genome (Irwin et al. 2018; Toews et al. 2016). This pattern suggests differential gene flow throughout the genome and incomplete reproductive isolation. The exact reproductive barriers are unknown, but are likely post-zygotic, given the lack of assortative mating (Brelsford & Irwin 2009). These characteristics - a hybrid zone of intermediate age, strong but incomplete reproductive isolation, large population size, and high dispersal - make the yellow-rumped warbler hybrid zone well suited to an investigation of the genetic basis of reproductive barriers. These characteristics ameliorate the effects of drift on local populations, enabling the use of cline analysis to study patterns of gene flow.

In this study, we investigate the genetic architecture of reproductive barriers between *Setophaga coronata coronata* and *S. c. auduboni* and examine our findings in the context of theoretical predictions regarding the distribution and associations of loci involved in reproductive barriers. This analysis depends on accurate identification of the genes or genomic regions involved in reproductive barriers, so we utilize cline analysis, a well-established method for identifying regions where gene flow is restricted, and explore novel, alternative methods for detecting symmetric and asymmetric barriers to introgression at the edges of a hybrid zone. We leverage a large number of individual samples spanning the hybrid continuum and whole genome sequence data to provide

nearly ideal conditions for the identification of reproductive barrier loci. Lastly, we examine cline shape, cline coincidence, and patterns of linkage disequilibria among our putative barrier loci, contrasted against putatively neutral loci, to determine how genetic associations among reproductive barriers influence the total barrier to gene flow.

Methods

Data collection and preparation

This study uses the samples collected and sequence data generated as described in Chapter 1 (see Chapter 1 Methods for details). Briefly, we captured yellow-rumped along five transects that cross the hybrid zone and in allopatric populations. All captured birds were banded and released at the site of capture following acquisition of a sample of blood, morphological measurements, and photographs of plumage characteristics. We extracted DNA from blood samples using standard procedures for the Qiagen DNEasy blood and tissue kit, then prepared whole-genome libraries using the Seqwell Plexwell 384 LP kit, and sequenced on the NovaSeq 6000 for a target depth of 7x per bird. We included sequence data from additional samples collected in New York (Baiz et al. 2021) and New Mexico, accessed through the NCBI Sequence Read Archive (Szarmach, Brelsford, Witt & Toews 2021) to bolster our sampling of allopatric *coronata* and *auduboni* populations. To prepare the sequence data for analysis, we applied per-site and per-individual missingness and depth filters and pruned our data of close relatives (see Chapter 1 Methods for details).

Cline analysis

Fitting clines to SNP data is only effective for sites with high inter-population differentiation. For cline fitting, we filtered our data for SNPs that have a difference in allele frequency of at least 0.8 between allopatric populations of *coronata* and *auduboni*. We determined that a population is allopatric if the average admixture-based hybrid index is at most 0.10 for *auduboni* or at least 0.90 for *coronata*. We excluded from cline analysis 2 individuals captured outside the breeding period (April through July), and 6 individuals that were not breeding at the time of capture, indicated by the lack of territorial behavior exhibited by the birds (i.e. birds on migration can sometimes be captured but will not sing or remain in the area of capture a day later) or the presence of excess body fat. Because the cline analysis also requires accurate information about the distance of each locality from the midpoint of the hybrid zone, we also excluded an additional 55 birds from localities where the distance was uncertain. We then computed the allele frequency for each of the highly differentiated SNPs at each remaining locality.

Cline fitting was performed using HZAR v0.2-5 (Derryberry, Derryberry, Maley & Brumfield, 2014). To obtain a cline for genome-wide ancestry as our point of comparison, we averaged ancestry estimates computed in Admixture (Alexander, Novembre & Lange 2009) for each locality, and fit clines to these data using 3 different models that differ in cline tail fitting. For the cline in genome-wide ancestry, we used the quantitative trait model in HZAR, and we fit model variations that estimate mirrored tails, both tails separately, or fit a sigmoid cline with no tail parameters estimated. We compared the AICc scores for each model and selected for our analysis the one with the

lowest score. Although there is some slight variation in genome-wide ancestry clines among transects across different regions of the hybrid zone and at least one transect that varies significantly over time (see Chapter 1), we chose to aggregate all localities into a single transect for cline fitting. Combining transects increases sample size and ameliorates the effects of genetic drift, so should provide a more robust estimation of ancestry or allele frequencies at any distance from the hybrid zone and improve cline fitting accuracy. For individual SNP clines, we fit variations of the biallelic genetic locus model, that use either the empirical allele frequency or fixed frequency at 0 or 1 at the two most distant sites from the hybrid zone midpoint and estimate mirrored tails or both tails separately. After fitting models to each highly differentiated SNP, we selected data from the best performing model, determined from a comparison of AICc scores as above.

HZAR estimates several parameters that describe the shape and position of a cline (Derryberry, Derryberry, Maley & Brumfield 2014; Gay, Crochet, Bell & Lenormand 2008). We will focus on three of these parameters. Cline center describes the point in space along a transect where the ancestry of a hypothetical population is 0.5, relative to parental populations and is measured in units of distance (km). Cline width is defined by the inverse of the maximum slope of the central portion of the cline, and describes the distance over the transect where the greatest change in allele frequency, or other trait value, occurs. Last, we estimate the parameter τ which is the coefficient for the slope of the tails of the cline. These tails are estimated by HZAR independently in the "both" model described above, or mirrored, as specified by the user.

Binomial test for introgression

Geographic cline analysis can only be applied to genomic regions that are highly differentiated between populations. As a complement to the cline analysis, we tested for introgression beyond the edges of the hybrid zone by comparing observed and expected allele frequencies for two hybrid classes: birds with ancestry between 0.1 and 0.2 (*auduboni*-like), and birds with ancestry between 0.8 and 0.9 (*coronata*-like). A two-tailed binomial test was performed for each SNP with an allele frequency difference between allopatric populations of at least 0.3, taking the reference allele frequency as the observed frequency and the average genome-wide ancestry for the corresponding hybrid class multiplied by the frequency of the reference allele in allopatric *coronata* as the expected frequency. Our chosen significance threshold for a deviation from neutral introgression is 1/294877, the reciprocal of the number of SNPs in the analysis. This significance threshold is equivalent to a false-discovery rate of one for the analyzed dataset, which is less conservative than a Bonferroni correction while still providing a stringent threshold for significance. Restricted introgression is indicated when the observed reference allele frequency is closer to 0 than expected in *auduboni*-like birds and closer to 1 than expected in *coronata*-like birds. The reverse indicates excess introgression.

Binning SNPs by differentiation peaks

We assigned SNPs to genomic intervals to facilitate our analysis of cline parameters and genetic associations. The assignment of SNPs to intervals serves several purposes. Primarily, it allows us to summarize parameter estimates across a genomic

region to ameliorate noise in individual genotypes associated with low-coverage sequence data. Computation of pairwise linkage disequilibria (more properly, correlation) is more manageable when performed over a small number of genomic intervals rather than tens of thousands of SNPs. Summarizing parameter estimates over these intervals is valid if recombination within intervals is low. We defined these intervals as spans of SNPs where allele frequency differences between allopatric populations is greater than 0.8. If a highly differentiated SNP is within 1 Mbp of another highly differentiated SNP, those two SNPs and all intervening SNPs are assigned to the same interval. For each interval, we computed principal components using PLINK v1.90b6.25 (Purcell et al. 2007) and plotted eigenvectors of principal components 1 and 2 to visually identify genotype clusters that represent homozygotes for the reference and non-reference alleles, and heterozygotes. Intervals for which principal component 1 did not separate parental and hybrid genotypes (e.g. instead parsed individuals by sex) were excluded from downstream analyses. An example of our use of principal components to evaluate patterns of genotypic clustering within differentiation peaks is provided in Figure S2.1. Tight genotype clusters with few intermediates suggests that there is little recombination within the interval, since tightly linked adjacent alleles will behave somewhat like a single locus.

Genetic interactions

To assess the relative strength of linkage disequilibria among barrier loci and non-barrier loci, we classified a differentiation peak as a barrier locus if the average cline width for that region was narrower the genome-wide ancestry cline width. We estimated

the pairwise correlation coefficient of genotypes among pairs of differentiation peaks on different chromosomes in R v4.2.1 (R Core Team 2013). First, SNPs were assigned to intervals as outlined above. We then generated principal components for genotypes within each block using Plink v1.90b6.25 (Purcell et. al., 2007). We then computed the pairwise correlation coefficient of principal component 1 of each interval in a pair among all individuals that passed the relevant filtering steps. We then visually examined the distributions of correlations between pairs of barrier loci, non-barrier loci, and barrier and non-barrier loci for qualitative differences.

Identification of centromeric regions

Because the *S. coronata* reference genome was assembled entirely from short reads, highly repetitive genomic regions such as centromeres are likely to be missing from the assembly. We therefore searched for putative centromere-associated repetitive sequences in the long-read-based assembly of a related species, *Setophaga petechia* (Tsai et al. 2024). We identified repetitive sequence motifs between 50 and 400 base-pairs long using TRASH (Wlodzimierz, Hong & Henderson 2023). After identifying a highly abundant 286 – 287 bp sequence, we aligned *S. petechia* contigs containing this sequence to the *S. coronata* genome assembly. Specifically, we extracted 1 kbp of sequence from the *S. petechia* assembly at intervals of 10 kbp and aligned these 1 kbp sequences to the *S. coronata* assembly using bwa-mem2 (Vasimuddin et al. 2019). If at least 3 consecutive 1 kbp sequences containing the repeat motif aligned to the *S. coronata* reference assembly in the correct orientation and spaced 9 kbp apart, the

position of the alignment closest to the centromeric repeat in *S. petechia* was taken as the approximate putative centromere position.

Results

Data collection

After filtering for genotype quality and missing data, we retained 1208 individuals out of 1214 *coronata*, *auduboni*, and their hybrids. Six of the samples that passed our quality filters were excluded, since each was closely related to another individual in our data set, and another was excluded as it represented a likely mis-labeled duplicate of another bird; leaving 1201 total birds for general analysis. Alignment rate for the reads we obtained is, on average, approximately 84 %. Per-individual average sequence depth is approximately 6. After filtering, we retained a total of 14,709,110 SNPs.

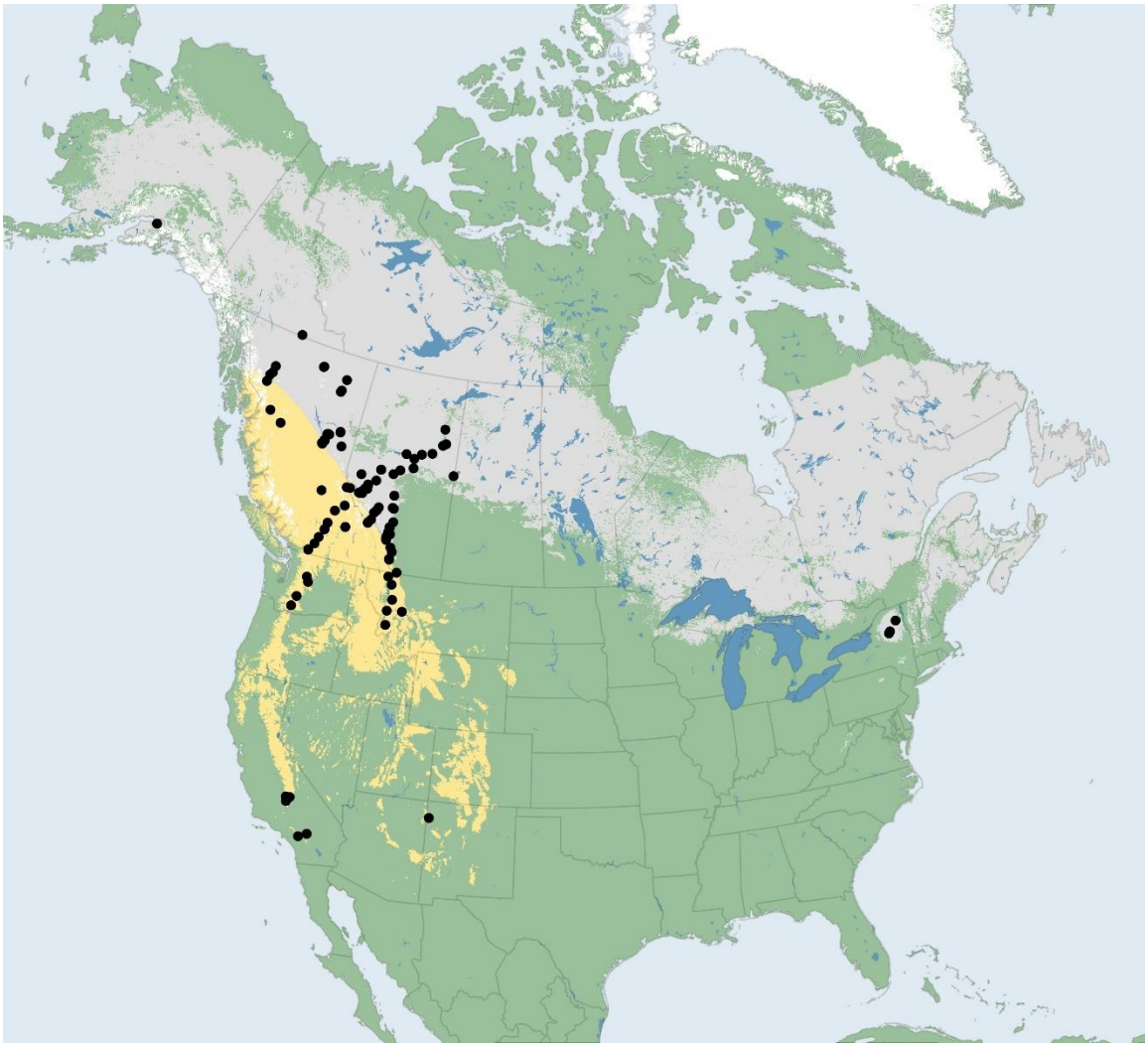

Figure 2.1: Localities sampled (black points) and the breeding ranges of *coronata* (gray) and *auduboni* (yellow). A summary of each locality is provided in Table S2.1.

Cline analysis

The genome-wide cline in ancestry (Figure 2.2) has a center at -2.7 km (-5.5 km, -1.2 km) and width of 170 km (158 km, 179 km). The model with both tails estimated separately performed best, with an AICc score 79 points lower than the mirrored tail model.

After filtering our SNP data for sites with an allele frequency difference greater than 0.8 between allopatric populations of the parental subspecies, we retained 41,014

SNPs to use for our per-SNP cline analysis. Individual and locality filtering retained 1138 birds.

The SNP clines vary across the genome but are relatively consistent within runs of SNPs on the same chromosome, albeit with some noise in the point estimates of all parameters (Figure 2.3). Given that cline parameters within consecutive runs of SNPs are consistent, we summarized cline parameters by taking the average of cline parameters for an interval of the genome defined by a peak of differentiation (as outlined in the methods section on binning by differentiation peaks). Binning SNPs by differentiation peaks results in 93 intervals. The first principal component of genotypes in these intervals separates individuals by ancestry for all but two. Interval 74 on chromosome 24 and interval 93 on the Z chromosome, both of which discriminate sex on principal component 1 and ancestry on principal component 2. Our results show that several regions of the genome exhibit narrower clines than the genome-wide ancestry cline (170 km), which we will refer to as our putative barrier loci (Figure 2.3B). Of, these barrier loci, 33% (14/42) are within chromosomes larger than 50 Mbp excluding the sex chromosomes, and 31% (13/42) are within the Z chromosome.

Cline center (Figure 2.3A) shows a significant negative relationship with cline width – intervals with narrow clines have more eastward-shifted centers (r^2 = -0.34, p-value = 0.00077). The mean cline center of our barrier loci is 2.9 km compared to -2.6 km for our non-barrier loci.

In addition to estimating the center and width parameters for the central step of each SNP cline, we estimated a tail parameter that describes where, in relation to the

cline center, the cline transitions from the central step to the exponential decay in the frequency of a foreign allele (δ) as well as a parameter for the coefficient of the slope of the tail of introgression (τ) (Derryberry, Derryberry, Maley & Brumfield 2014; Gay, Crochet, Bell & Lenormand 2008). The tail parameters estimated for each differentiated region broadly overlap and appear noisy (Figure 2.3C).

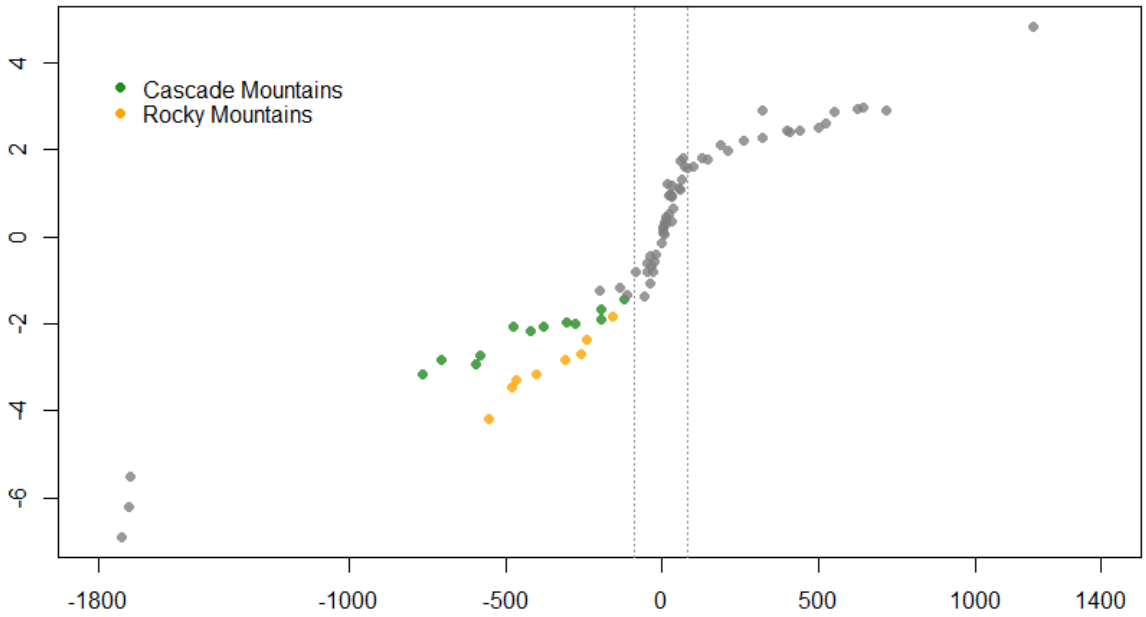

Figure 2.2: Logit-transformed cline of genome-wide ancestry. Points represent the average ancestry of each sampled locality. Dashed lines indicate the transition point from the central step and the shallower tails (-87 km and 83 km) for a cline width of 170 km.

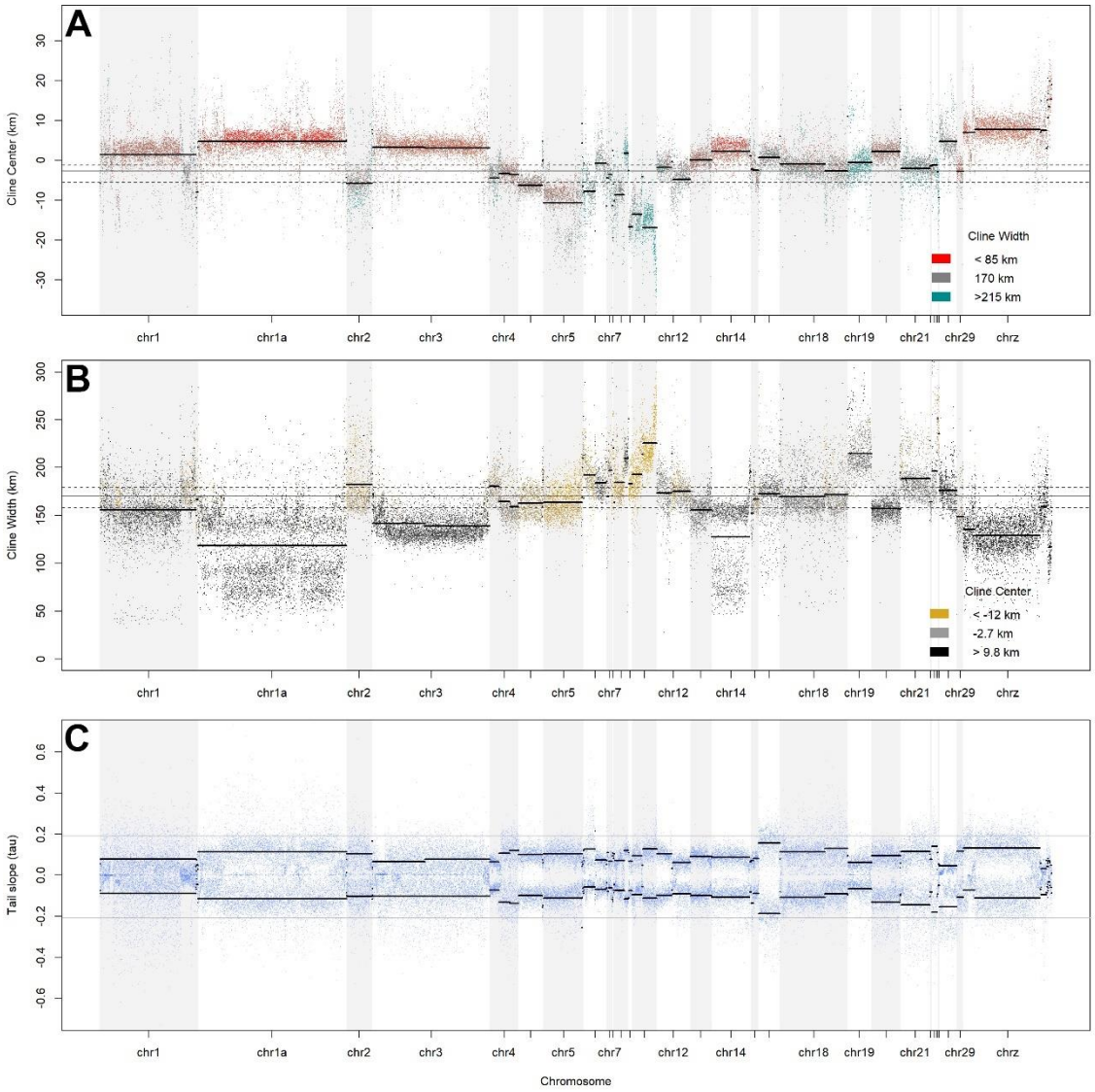

Figure 2.3: Summary of cline parameters estimated for 41,014 SNPs in highly differentiated regions. Cline center is plotted in panel A, with color corresponding to the width of the cline for each SNP. Panel B shows cline width for each SNP, with color corresponding to the cline center. In panels A and B, the genome-wide ancestry cline center or width is represented by the horizontal line across each plot bounded by confidence intervals indicated with a dashed line. Panel C shows the coefficients for the western (negative values) and eastern (positive values) cline tails for each SNP. Black horizontal bars represent the average center, width, or tail parameters for SNPs within each highly differentiated interval. X-axis for each plot is scaled so that the distance between SNPs is equal, and does not reflect the actual chromosomal position. This scaling is needed to show variation within intervals that would be obscured given the narrowness of differentiated regions relative to whole chromosomes.

Binomial test for introgression

We identified 19816 SNPs with restricted introgression and 2395 SNPs with excess introgression into *auduboni*-like hybrids (hybrid index between 0.1 and 0.2; n = 180). For *coronata*-like hybrids (hybrid index between 0.8 and 0.9; n = 208), we found 5122 SNPs with restricted introgression, and 1006 SNPs with excess introgression. In summary, there is a strong signal of asymmetric introgression of *coronata* alleles into the *auduboni*-like hybrids, especially among the large autosomes (1, 1a, and 3) and the Z chromosome. The results in *coronata*-like hybrids are noisier and show that regions with restricted introgression are located adjacent to regions where introgression is in excess (Figure 2.4).

Genetic associations

We computed the pairwise correlation coefficient of principal component 1 genotypes within peaks of differentiation. Plotting principal components 1 and 2 for each peak generally shows discrete genotype clusters representing homozygotes for the reference and non-reference alleles, and an intermediate cluster representing heterozygotes. The formation of discrete clusters suggests that recombination within these regions is limited (example provided in Figure S2.1), so we averaged cline parameter estimates within peaks to ameliorate noise in individual SNP parameter estimates and simplify downstream analysis. These intervals defined by peaks of differentiation were categorized as “barrier loci” on the condition that the mean of cline width for SNPs within an interval is less than the genome-wide ancestry cline width of 170 km.

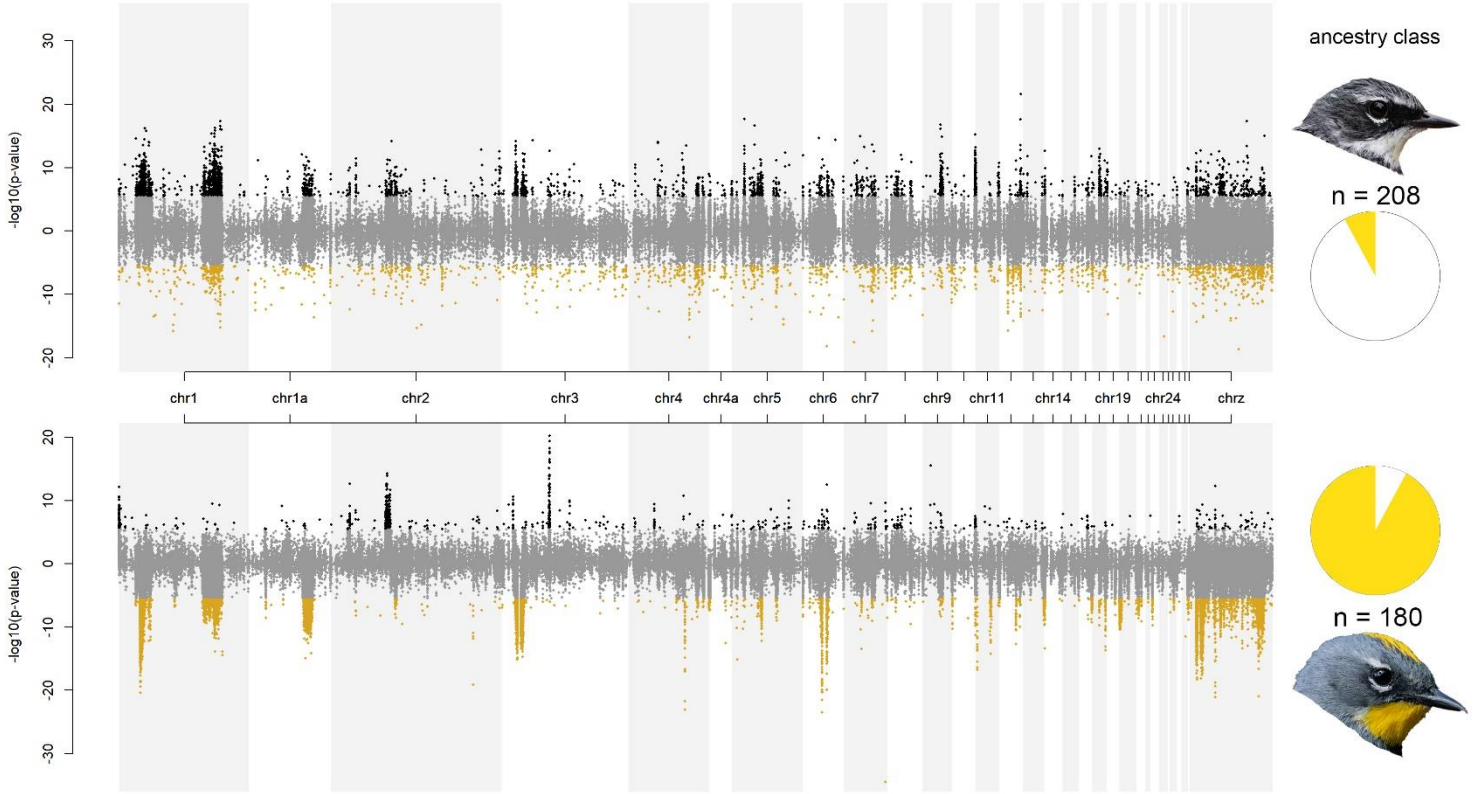

Figure 2.4: Binomial test for introgression. Colored points represent SNPs with significantly more *coronata*-like (black) or *auduboni*-like (yellow) allele frequency than expected in hybrid classes where genome-wide ancestry is 80 - 90% *coronata* (top) and 10-20% *coronata* (bottom). Introgression is considered "restricted" if alleles of the minority-ancestry subspecies is present at a lower frequency than expected.

Pairs of barrier loci exhibit stronger correlations (mean = 0.701, s = 0.324) than do pairs of non-barrier loci (mean = 0.657, s = 0.397) or pairs of non-barrier and barrier loci (mean = 0.679, s = 0.360). A permutation test, in which we randomized categorization of each locus in a pair as “barrier” or “non-barrier”, shows that the observed median genotypic correlation between non-barrier locus pairs is below the median of randomized pairs, and that the observed median genotypic correlation between barrier locus pairs is greater than permuted pairs (Figure 2.5D). We find that pairs of non-barrier loci tend to have much more coincident clines (i.e. similar center) than random pairs of loci, and that pairs of barrier loci tend to differ slightly but significantly in their cline center (p-value = 0.013; Figure 2.5E).

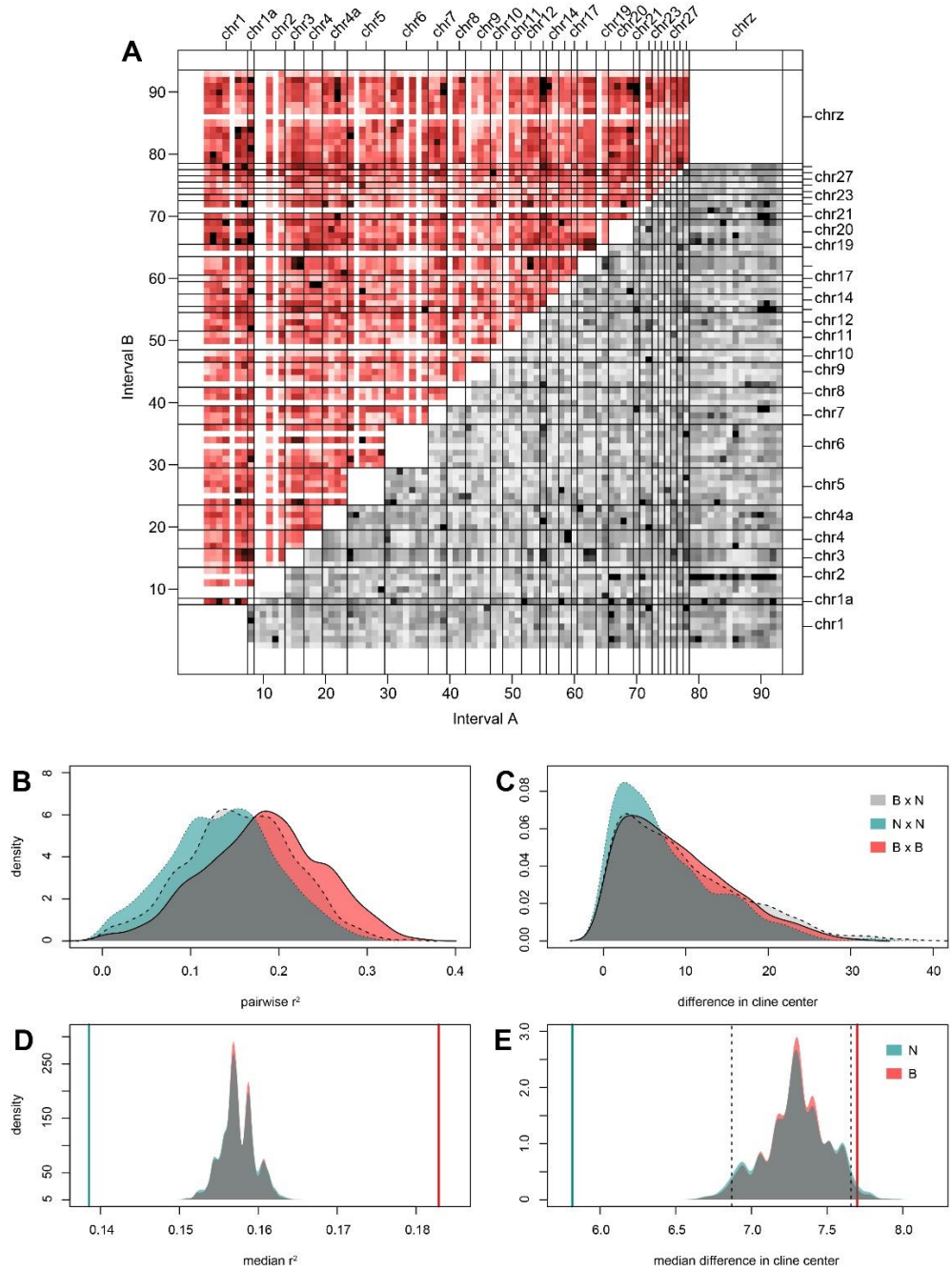

Figure 2.5: The heatmap (A) shows r^2 of genotype clusters inferred from PCA for each interval (red) and mean r^2 of the highly differentiated SNPs within intervals (gray). Darker colors correspond with stronger correlations and black points represent the top 5%. Panels B and C represent the distribution of genotype cluster r^2 (panel B) or the difference in cline center (panel C) of intervals categorized as "barrier" - B or "non-barrier" – N, based on the average cline width of loci in the interval. Panels D and E show results from a permutation test comparing the median r^2 (D) or difference in cline center (E) of locus categories.

Chromosome A	Interval A	Chromosome B	Interval B	r^2
chr1a	8	chr20	66	0.358423
chr14	56	chrz	91	0.352804
chr21	70	chrz	91	0.347681
chr21	70	chrz	90	0.346512
chr1a	8	chrz	92	0.338861
chr1a	8	chr12	52	0.33749
chr4a	22	chrz	90	0.329839
chr5	26	chr15	58	0.323994
chr15	58	chr20	69	0.323752
chr13	55	chrz	90	0.322573
chr13	55	chrz	91	0.321742
chr1a	8	chr15	58	0.316577
chr5	24	chrz	83	0.316211
chr5	26	chr6	34	0.314019
chr1	7	chr22	72	0.314006
chr1	2	chr20	66	0.312455
chr1a	8	chr20	67	0.312134
chr12	53	chr18	62	0.310835
chr1	4	chr29	78	0.30911
chr4	18	chr15	59	0.307378
chr20	69	chrz	84	0.30585
chr4	19	chr15	59	0.304775
chr13	55	chrz	92	0.303799
chr1	4	chr20	69	0.303736
chr1	2	chr20	67	0.303193
chr3	16	chr18	62	0.30294
chr1a	8	chrz	81	0.302919
chr10	47	chr28	77	0.302254
chr4a	22	chrz	89	0.302162
chr1	2	chrz	80	0.301133

Table 2.1: Top 1% pairwise correlations between principal component 1 of genotypes within peaks of differentiation (see Figure 2.5A, red heatmap). Each peak of differentiation within a chromosome was given a unique numerical identifier, shown here as a number in the “Interval A” and “Interval B” columns.

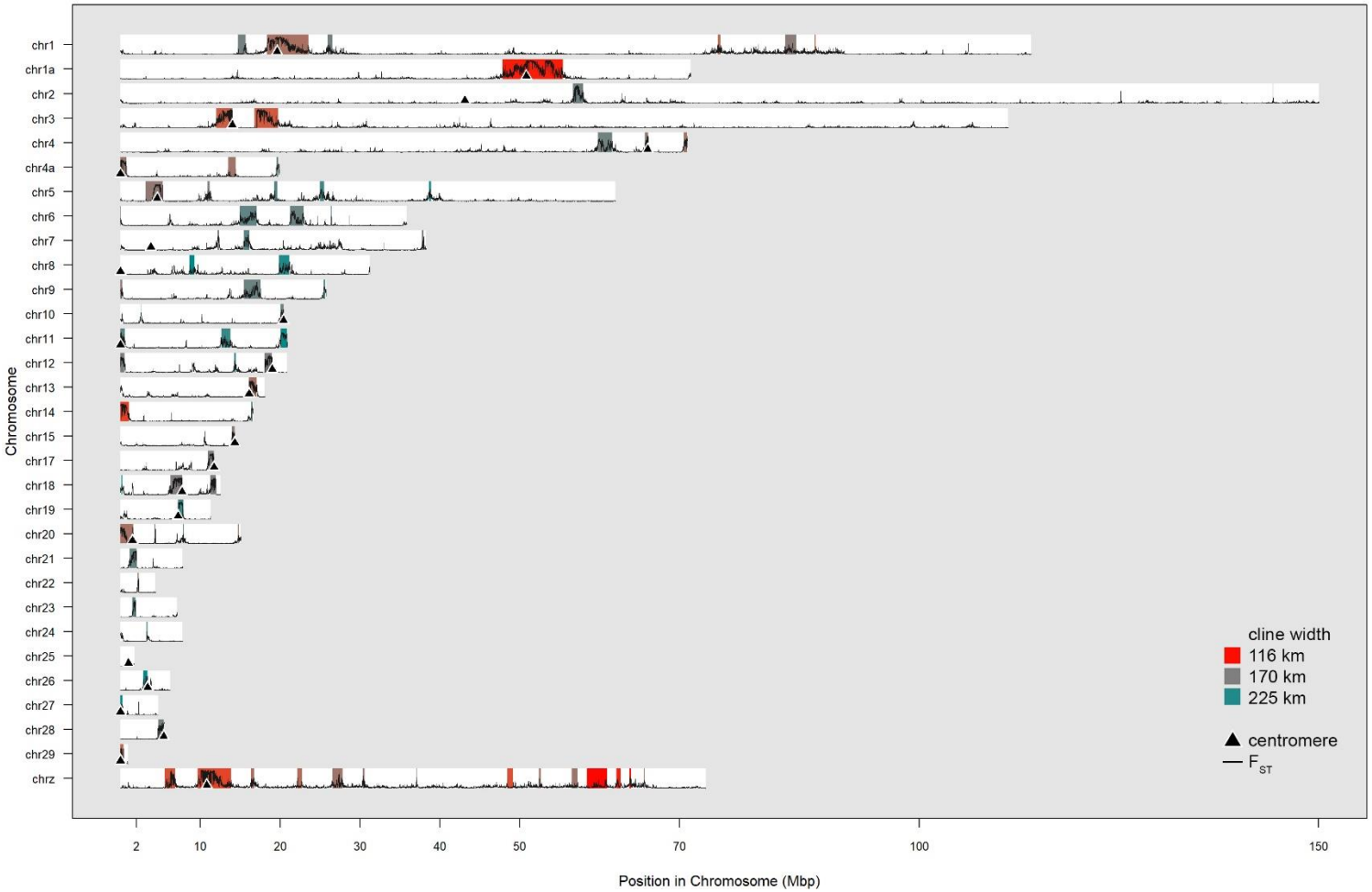

Figure 2.6: Overview of cline width, differentiation, and centromere positions in the *Setophaga coronata coronata* genome assembly. Color represents the average cline width of SNPs within each interval defined by a peak of differentiation. Weighted F_{ST} in 10 kbp windows, plotted within each chromosome as a black line, is scaled to the height of each chromosome, from 0 at the bottom and 1 at the top. Triangles represent the location of putative centromere-associated sequence motifs.

Discussion

The genetic architecture of reproductive isolation

We find that the extent and symmetry of introgression is heterogeneous across the genome. Putative barrier loci are clustered on large chromosomes, especially the Z chromosome. The Z chromosome makes up only about 7% of the genome yet it harbors 31% of the intervals we categorized as putative barrier loci. Barrier loci are over-represented on the Z chromosome, suggesting that the distribution of "speciation genes" may be non-random and that the sex chromosomes are important components of reproductive barriers. Overall, our results suggest that reproductive isolation between *Setophaga coronata coronata* and *S. c. auduboni* is driven by many genetic variants, each with small effects, and that neither variation in recombination rate nor differentiation fully account for the presence of barrier loci. Recombination rate tends to be lower near centromeres and within larger chromosomes (Nachman 2002), so if low recombination rate is particularly important for the maintenance of divergent genomes, we may expect an over-representation of genome regions experiencing a strong barrier to gene flow within the large autosomes. Instead, relative to their total contribution to the genome, we find fewer barrier loci than expected (33%) on the autosomes larger than 50 Mbp even though these chromosomes make up about 57% of the genome. This finding contrasts the initial interpretation of peaks of differentiation as "islands of speciation" (Turner, Hahn, & Nuzhdin 2005) as well as the counterargument that neutral divergence accumulates in regions of restricted recombination (Noor & Bennett 2009) and aligns with more recent observations in anurans, for example (Dufresnes et al. 2021). The large proportion of

barrier loci on the Z chromosome provides further support for the observation that sex chromosomes, relative to the autosomes, have a disproportionately large contribution to reproductive barriers (Presgraves 2008; Coyne 1992; Coyne 1989) as they can have a strong influence on traits that influence viability or sterility of hybrids.

Multilocus cline theory (Barton & Gale 1993) and more recently developed theories (Flaxman, Wacholder, Feder, & Nosil 2014) indicate that the genetic architecture of speciation with gene flow involves strong associations among barrier loci. We find evidence that strong linkage disequilibria among our putative barrier loci are maintained despite extensive hybridization, providing empirical support for the role of the "coupling" of barrier effects (Dopman, Servedio, Butlin & Smadja 2024) in the process of speciation.

Cline analysis

We find that clines among our putative barrier loci are shifted eastward relative to the genome-wide ancestry cline. This pattern is consistent with two possibilities. Overall, gene flow may be asymmetric due to differential dispersal tendency or unbalanced population sizes. Greater long-distance dispersal of *coronata* relative to *auduboni*, or a larger population of *coronata* could shift the balance of gene flow and lead to a greater influx of neutral *coronata* alleles into the hybrid zone. Loci under selection will experience restricted introgression and their clines will remain stationary while the neutral regions of the genome continue to introgress asymmetrically. It is unlikely that the difference in cline center between barrier and non-barrier loci is due to positive selection favoring *coronata* alleles, because the clines for beneficial alleles should be wide, and any universally beneficial allele would likely have swept to fixation within the

time since secondary contact was established. Alternatively, the difference in cline center between barrier loci and neutral peaks of differentiation could be explained by stronger selection against *coronata* alleles at barrier loci that are introgressed into an *auduboni* background than the reverse. Asymmetric dispersal and asymmetric selection are not mutually exclusive and may work together to produce asymmetric gene flow.

We aimed to use information from the tails of the clines to improve the resolution of our scan for barrier loci. The slope of the tail of a geographic cline corresponds with selection on an individual locus, as it represents the decay of the frequency of introgressing heterospecific alleles after generations of backcrossing and recombination. Despite thorough sampling of the edges of the hybrid zone, we are unable to analyze the tails of the clines due to high levels of noise and low variation in the estimated tail parameters. We recommend that researchers drawing inferences from the tails of geographic clines should do so with caution. Alternative methods for quantifying patterns of gene flow at the edges of the hybrid zone, such as the method discussed below, may be more effective.

Introgression beyond the hybrid zone

The strongest pattern we observed from our analysis of introgression in nearly pure *coronata* and *auduboni* populations is that there is broad asymmetry, particularly among the large autosomes and the Z chromosome. *Coronata* alleles appear to be strongly disfavored in the *auduboni* genetic background. This pattern is consistent with our results from the cline analysis that show a strong displacement of the cline center of Z chromosome loci, with a similar pattern on chromosomes 1a and 3 (Figures 2.3 & 2.6).

We find that our analysis of introgression into *coronata*-like hybrids is noisier than in *auduboni*-like hybrids. This difference may be due to differences in ancestral polymorphism in the two populations. While our sampling of allopatric *auduboni* populations likely includes most of the genetic variation present, our sampling of allopatric *coronata* populations is limited in the eastern extent of their breeding range. The wide geographic range of *coronata*, which breeds in territories extending from Alaska to the north-eastern coast of the United States and Canada, may include greater genetic diversity than in *auduboni*. Because our sampling of the eastern extent of the *coronata* breeding range is limited to only 5 individuals and we have no representatives of the range between the provinces of Saskatchewan and Ontario, we may not have captured some of the polymorphism that exists in *coronata* broadly. Failing to capture this variation may have led us to erroneously interpret a signal of ancestral polymorphism in allopatric *coronata* as a signal of excess introgression of *auduboni* alleles.

Our test for introgression at the edges of the hybrid zone provides higher resolution throughout the genome than our cline analysis. It also offers additional information, as it can detect restricted or excess introgression. The presence of some narrow regions where introgression is highly restricted provides candidate regions to explore for individual genes that may contribute to reproductive barriers.

Although our analysis of the cline tails was not able to achieve our intended results, the test of differential introgression serves a similar purpose, in that we are able to identify narrow genomic regions experiencing selection against introgression in nearly pure populations, where alleles from one population are isolated in the opposing

population's genetic background. One advantage of this analysis is that it is not limited to highly differentiated or fixed differences between parent populations. In the early stages of speciation with gene flow, weakly selected loci that contribute to reproductive barriers may persist at intermediate frequency for many generations. Therefore, limiting a scan to highly differentiated regions may provide an incomplete characterization of the genomic regions involved in reproductive isolation.

Genetic interactions

Theory predicts that speciation with gene flow is possible when divergent selection maintains linkage disequilibria among loci involved in reproductive barriers (Flaxman, Wacholder, Feder, & Nosil 2014) and that linkage disequilibria among barrier loci can increase the total strength of selection against hybrids, generating a stronger barrier to gene flow than selection on an individual locus (Barton & Gale 1993). Consistent with these predictions, we find that genotypic associations among our putative barrier loci are strong, and the "stepped" shape of the genome-wide ancestry cline and clines fit to individual SNPs suggests genome-wide coupling of barrier effects that increase the total selection against hybrids (Figure 2.2). Additionally, we find that genotypic associations are stronger among loci with narrow clines (Figure 2.5). In particular, differentiation peaks on the Z chromosome tend to exhibit strong linkage disequilibria, featuring prominently in the strongest 1% of correlations (Table 2.1). However, cline center is not coincident among barrier loci. This pattern seems to be driven by the strong asymmetry in introgression we observe among loci on the Z

chromosome, which comprise 31% of our putative barrier loci. In contrast, the center of clines estimated for non-barrier loci are highly coincident (Figure 2.5E).

The strong associations among unlinked barrier loci may include loci involved in genetic incompatibilities. Multi-locus incompatibilities are thought to be an important factor in generating and maintaining post-zygotic reproductive isolation after secondary contact (Dobzhansky 1937; Muller 1942). Our analysis of genotype correlations among unlinked loci reveal a small number of outliers that are good candidates for further analysis to determine whether these exhibit signatures consistent with negative epistasis.

Genic incompatibilities involving the sex chromosomes impose barriers to gene flow

The importance of heteromorphic sex chromosomes in the evolution of post-zygotic intrinsic reproductive barriers is well supported by empirical examples. Interactions involving sex chromosomes commonly result in reduced viability or sterility (e.g. Presgraves 2003; Orr & Irving 2001; Orr 1987). Our cline analysis and test for asymmetric introgression both show that the sex chromosomes exhibit a strong barrier to gene flow, and that the barrier is asymmetric. We also find that linkage disequilibrium between pairs of putative barrier loci is higher than between pairs of non-barrier loci and the Z chromosome is commonly found among locus pairs with the strongest linkage disequilibria. This pattern suggests that in yellow-rumped warblers, between-locus incompatibilities involving the sex chromosomes may play a major role in reproductive isolation. Closer examination of putative barrier loci focused on identifying signatures of asymmetric negative epistasis will clarify the role of genetic incompatibilities in this system. This work demonstrates that large scale genome-scans have the power to

elucidate the genome-level signatures and genic components of reproductive barriers and advance our understanding of the genetic architecture of speciation with gene flow.

References

Alexander, D.H., Novembre, J., and Lange., K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 19:1655–1664.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of molecular biology, 215(3), 403-410.

Andrew, R. L., & Rieseberg, L. H. (2013). Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. Evolution, 67(9), 2468-2482.

Baiz, M. D., Wood, A. W., Brelsford, A., Lovette, I. J., & Toews, D. P. (2021). Pigmentation genes show evidence of repeated divergence and multiple bouts of introgression in Setophaga warblers. Current Biology, 31(3), 643-649.

Barton, N. H. (1979). Gene flow past a cline. Heredity, 43(3), 333-339.

Barton, N. H., & De Cara, M. A. R. (2009). The evolution of strong reproductive isolation. Evolution, 63(5), 1171-1190.

Barton, N. H., Gale, K. S., & Harrison, R. G. (1993). Genetic analysis of hybrid zones. Hybrid zones and the evolutionary process, 13-45.

Barton, N. H., & Hewitt, G. M. (1985). Analysis of hybrid zones. Annual review of Ecology and Systematics, 16(1), 113-148.

Brelsford, A., & Irwin, D. E. (2009). Incipient speciation despite little assortative mating: the yellow-rumped warbler hybrid zone. Evolution, 63(12), 3050-3060.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., ... & Ellegren, H. (2015). Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of Ficedula flycatchers. Genome research, 25(11), 1656-1665.

Coyne, J. A. (1989). Two rules of speciation. Speciation and its consequences, 180-207.

Coyne, J. A. (1992). Genetics and speciation. Nature, 355(6360), 511-515.

Coyne, J. A., & H Allen Orr. (2004). Speciation. Sinauer Associates, Oxford University Press.

Carneiro, M., Ferrand, N., & Nachman, M. W. (2009). Recombination and speciation: loci near centromeres are more differentiated than loci near telomeres between subspecies of the European rabbit (Oryctolagus cuniculus). Genetics, 181(2), 593-606.

Derryberry, E. P., Derryberry, G. E., Maley, J. M., & Brumfield, R. T. (2014). HZAR: hybrid zone analysis using an R software package. Molecular ecology resources, 14(3), 652-663.

Dobzhansky, T. (1937). Genetics and the Origin of Species. Columbia university press.

Dopman, E. B., Shaw, K. L., Servedio, M. R., Butlin, R. K., & Smadja, C. M. (2024). Coupling of barriers to gene exchange: causes and consequences. Cold Spring Harbor Perspectives in Biology, a041432.

Dufresnes, C., Brelsford, A., Jeffries, D. L., Mazepa, G., Suchan, T., Canestrelli, D., ... & Crochet, P. A. (2021). Mass of genes rather than master genes underlie the genomic architecture of amphibian speciation. Proceedings of the National Academy of Sciences, 118(36), e2103963118.

Feder, J. L., Gejji, R., Yeaman, S., & Nosil, P. (2012). Establishment of new mutations under divergence and genome hitchhiking. Philosophical Transactions of the Royal Society B: Biological Sciences, 367(1587), 461-474.

Feder, J. L., Nosil, P., & Flaxman, S. M. (2014). Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. Frontiers in Genetics, 5, 89179.

Felsenstein, J. (1981). Skepticism towards Santa Rosalia, or why are there so few kinds of animals?. Evolution, 124-138.

Flaxman, S. M., Feder, J. L., & Nosil, P. (2013). Genetic hitchhiking and the dynamic buildup of genomic divergence during speciation with gene flow. Evolution, 67(9), 2577-2591.

Flaxman, S. M., Wacholder, A. C., Feder, J. L., & Nosil, P. (2014). Theoretical models of the influence of genomic architecture on the dynamics of speciation. Molecular ecology, 23(16), 4074-4088.

Gay, L., Crochet, P. A., Bell, D. A., & Lenormand, T. (2008). Comparing clines on molecular and phenotypic traits in hybrid zones: a window on tension zone models. Evolution, 62(11), 2789-2806.

Harrison, R. G., & Larson, E. L. (2014). Hybridization, introgression, and the nature of species boundaries. Journal of Heredity, 105(S1), 795-809.

Huang, K., & Rieseberg, L. H. (2020). Frequency, origins, and evolutionary role of chromosomal inversions in plants. Frontiers in plant science, 11, 518896.

Hubbard, J. P. (1969). The relationships and evolution of the Dendroica coronata complex. The Auk, 86(3), 393-432.

Irwin, D. E. (2018). Sex chromosomes and speciation in birds and other ZW systems. *Molecular Ecology*, *27*(19), 3831-3851.

Jofre, G. I., & Rosenthal, G. G. (2021). A narrow window for geographic cline analysis using genomic data: Effects of age, drift, and migration on error rates. Molecular Ecology Resources, 21(7), 2278-2287.

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., ... & Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. Nature, 484(7392), 55-61.

Mila, B., Smith, T. B., & Wayne, R. K. (2007). Speciation and rapid phenotypic differentiation in the yellow-rumped warbler Dendroica coronata complex. Molecular Ecology, 16(1), 159-173.

Muller, H. J. (1942). Isolating mechanisms, evolution, and temperature. In Biol. Symp. (Vol. 6, p. 71).

Nachman, M. W. (2002). Variation in recombination rate across the genome: evidence and implications. *Current opinion in genetics & development*, *12*(6), 657-663.

Noor, M. A., & Bennett, S. M. (2009). Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. Heredity, 103(6), 439-444.

Orr, H. A. (1987). Genetics of male and female sterility in hybrids of Drosophila pseudoobscura and D. persimilis. *Genetics*, *116*(4), 555-563.

Orr, H. A., & Irving, S. (2001). Complex epistasis and the genetic basis of hybrid sterility in the Drosophila pseudoobscura Bogota-USA hybridization. *Genetics*, *158*(3), 1089-1100.

Payseur, B. A. (2010). Using differential introgression in hybrid zones to identify genomic regions involved in speciation. Molecular ecology resources, 10(5), 806-820.

Presgraves, D. C. (2003). A fine-scale genetic analysis of hybrid incompatibilities in Drosophila. *Genetics*, *163*(3), 955-972.

Presgraves, D. C. (2008). Sex chromosomes and speciation in Drosophila. Trends in Genetics, 24(7), 336-343.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics, 81(3), 559– 575.

R Core Team, R. (2013). R: A language and environment for statistical computing.

Szarmach, S. J., Brelsford, A., Witt, C. C., & Toews, D. P. (2021). Comparing divergence landscapes from reduced-representation and whole genome resequencing in the yellow-rumped warbler (*Setophaga coronata*) species complex. Molecular ecology, 30(23), 5994-6005.

Toews, D. P., Brelsford, A., Grossen, C., Milá, B., & Irwin, D. E. (2016). Genomic variation across the Yellow-rumped Warbler species complex. The Auk: Ornithological Advances, 133(4), 698-717.

Tsai, W. L., Escalona, M., Garrett, K. L., Terrill, R. S., Sahasrabudhe, R., Nguyen, O., ... & Bay, R. A. (2024). A highly contiguous genome assembly for the Yellow Warbler (Setophaga petechia). *Journal of Heredity*, *115*(3), 317-325.

Turner, T. L., Hahn, M. W., & Nuzhdin, S. V. (2005). Genomic islands of speciation in Anopheles gambiae. PLoS biology, 3(9), e285.

Turner, T. L., & Hahn, M. W. (2010). Genomic islands of speciation or genomic islands and speciation?. Molecular ecology, 19(5), 848-850.

Vasimuddin, M., Misra, S., Li, H., & Aluru, S. (2019, May). Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In *2019 IEEE international parallel and distributed processing symposium (IPDPS)* (pp. 314-324). IEEE.

Wlodzimierz, P., Hong, M., & Henderson, I. R. (2023). TRASH: tandem repeat annotation and structural hierarchy. *Bioinformatics*, *39*(5), btad308.

Wu, C. I. (2001). The genic view of the process of speciation. Journal of evolutionary biology, 14(6), 851-865.

Yeaman, S., & Whitlock, M. C. (2011). The genetic architecture of adaptation under migration–selection balance. Evolution, 65(7), 1897-1911.

Locality	State/Province	Distance	n	Hybrid Index
Abbott Bay	Montana	-310	9	0.06
Abraham Lake	Alberta	58	24	0.75
Alta Sierra	California	-1727	12	0.00
Anchorage	Alaska	1182	26	0.99
Beaver Creek	Montana	-481	9	0.03
Blaeberry	British Columbia	-30	15	0.31
Blue River	British Columbia	-123	10	0.19
Bob Quinn	British Columbia	-37	10	0.39
Britton Creek	British Columbia	-421	10	0.10
Chisholm	Alberta	397	10	0.92
Chuska Mtns	New Mexico	-1735	9	0.00
Clearwater	British Columbia	-195	8	0.16
Cline River	Alberta	36	32	0.66
Cold Lake	Alberta	523	5	0.93
Comstock	British Columbia	-377	10	0.11
Crowsnest	Alberta	-157	23	0.14
Etherington Creek	Alberta	-86	9	0.31
Evan Thomas	Alberta	-20	23	0.40
Fawcett Lake	Alberta	441	10	0.92
Fort McKay	Alberta	715	2	0.95
Fox Creek	Alberta	209	24	0.88
Gavin Lake	British Columbia	-200	17	0.23
Ghost	Alberta	32	9	0.71
Golden	British Columbia	-57	28	0.20
Grizzly Trail	Alberta	320	10	0.91
Hay Creek	Montana	-261	5	0.06
Hinton	Alberta	81	22	0.83
Hope	British Columbia	-476	22	0.11
House River	Alberta	549	9	0.95
Kennedy	British Columbia	-23	24	0.36
Lac LeJeune	British Columbia	-306	32	0.12
Livingstone	Alberta	-114	10	0.21
Lodgepole	Alberta	186	11	0.89
Maqua Lake	Alberta	642	9	0.95
McLeod	Alberta	262	10	0.90
Miette	Alberta	62	9	0.79
Mill Creek	Washington	-593	11	0.05
Moberly Lake	British Columbia	97	24	0.83
Morchuea Lake	British Columbia	65	12	0.86
Mt Adams	Washington	-766	9	0.04

Locality	State/Province	Distance	n	Hybrid Index
Mt Robson	British Columbia	-32	14	0.34
Nass River	British Columbia	-137	9	0.24
New York	New York	3113	5	1.00
Oconnor Lake	British Columbia	-277	10	0.12
Palisades	Alberta	29	15	0.59
Parker Meadow	California	-1702	19	0.00
Pelican Hills	Alberta	497	11	0.92
Peter Lougheed	Alberta	-40	26	0.25
Pine E	British Columbia	18	6	0.77
Pine Pass	British Columbia	22	18	0.72
Pine W	British Columbia	14	14	0.57
Pocahontas	Alberta	54	8	0.75
Pyramid Lake	Alberta	20	22	0.62
Rafter 6	Alberta	1	7	0.52
Rattlesnake	Montana	-466	3	0.04
Red Pass Junction	British Columbia	-48	11	0.31
Rocky Mountain House	Alberta	128	19	0.86
Saskatchewan RX	Alberta	12	43	0.59
Sherman Pass	California	-1697	5	0.00
Sibbald Creek	Alberta	11	19	0.61
Sibbald E	Alberta	9	18	0.58
Sibbald W	Alberta	-3	24	0.46
Skalkaho	Montana	-552	7	0.02
Slave Lake	Alberta	409	3	0.92
Smith	Montana	-402	10	0.04
Snaring	Alberta	28	9	0.72
South McMurray	Alberta	623	11	0.95
Spring Rimrock	Washington	-704	10	0.06
Sugarloaf	California	-1877	3	0.00
Sundance	Alberta	143	8	0.85
Sundre	Alberta	70	6	0.84
Todagin	British Columbia	27	29	0.73
Tumbler Ridge	British Columbia	58	11	0.85
Waiparous	Alberta	32	14	0.77
Waterton	Alberta	-243	7	0.09
Watson Lake	British Columbia	320	9	0.95
Whiskers Point	British Columbia	-48	23	0.36
White Pine	Washington	-579	11	0.06
Willow Creek	British Columbia	5	12	0.51
Yamnuska	Alberta	4	17	0.55
Yard Creek	British Columbia	-196	21	0.13

Locality	State/Province	Distance	n	Hybrid Index
Yellowhead Pass	British Columbia	3	18	0.54

Table S2.1: Summary of localities included in cline analysis for Chapter 2. Hybrid index is the average ancestry of birds sampled in each locality, relative to *coronata* ancestry, estimated with Admixture. Distance is the distance between the averaged coordinates for each bird's site of capture and the nearest point along the line joining the interpolated hybrid zone midpoint at each transect.

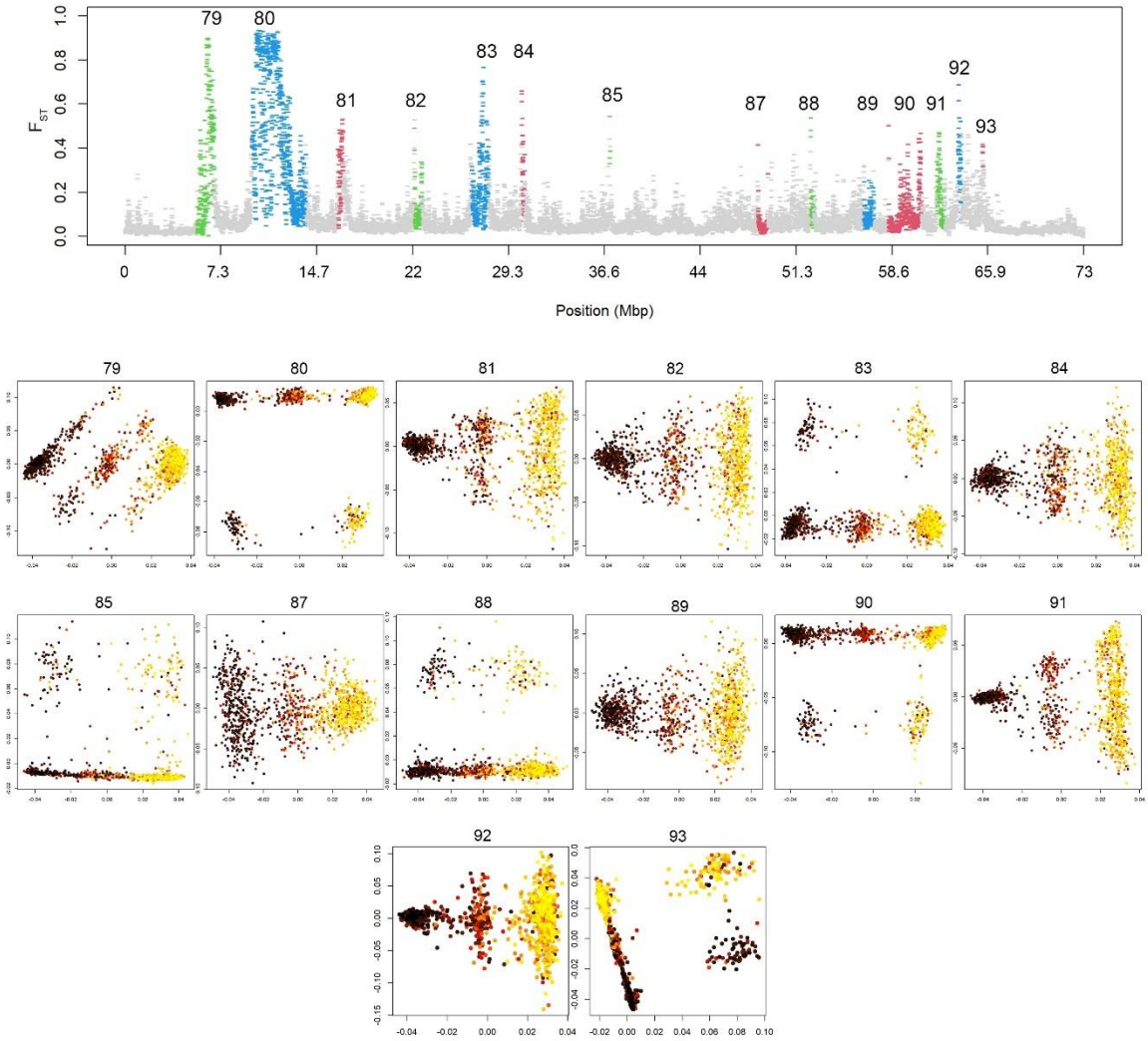

Figure S2.1: Top panel shows a Manhattan plot of differentiation peaks (weighted F_{ST} in 10 kbp windows), using the Z chromosome as an example. Peaks are colored to indicate intervals of consecutive SNPs with an allele frequency difference between allopatric *coronata* and *auduboni* > 0.8. Each peak is labeled above with its unique numerical ID. Below, scatterplots show each bird in the analysis, plotted with principal component 1 of genotypes within the interval on the x-axis and principal component 2 on the y-axis. Yellow points indicate *auduboni* ancestry, black points indicate *coronata* ancestry, and red points indicate a bird exhibits intermediate ancestry. Interval 86 is excluded, as it includes too few SNPs for principal components analysis. When three discrete genotype clusters align along the x-axis, the pattern suggests little recombination within the interval, where the clusters correspond with homozygotes for parental genotypes at either end and heterozygotes in the middle. Within interval 93, principal component 1 separates sex, and principal component 2 separates individuals by ancestry.

Chapter 3: Plumage genes are associated with reproductive barriers in hybridizing warblers

Background

Sexual selection may be an important factor in animal speciation as it can influence mating success within and between lineages and generate reproductive isolation when it results in divergence in sexual signals, mating success, or other aspects of pre-mating or pre-zygotic compatibility between populations (Mendelsson & Safran 2021; Panhuis, Butlin, Zuk, & Tregenza 2001). However, theoretical models exploring the conditions that allow speciation to occur by sexual selection provide restrictive conditions for sexual selection as the primary driver of speciation (reviewed in Turelli, Barton & Coyne 2001) and empirical studies often only indirectly link divergence in sexual traits with sexual selection and speciation (Coyne & Orr 2004; Maan & Seehausen 2011; Ritchie 2007; Panhuis, Butlin, Zuk & Tregenza 2001). Speciation involving sexual selection is most plausible when there is coordinated evolution of the genetic components of sexual signals, sensory systems that influence mate preference and choice, and divergent ecological selection, especially when populations interbreed (Servedio & Boughman 2017; Safran, Scordato, Symes & Rodriguez 2013).

Theoretical models of speciation with gene flow predict that coupling of the genetic components of reproductive barriers is needed to counteract genetic homogenization by migration and recombination (Flaxman, Wacholder, Feder, & Nosil 2014). Coupling can be accomplished through physical linkage or by selection that maintains favorable genotype combinations at unlinked loci. If speciation by sexual

selection is most likely when sexual signals and mate preferences are coupled, pleiotropy is perhaps the most likely genetic architecture underlying speciation driven by sexual selection (Xu & Shaw 2021; Naisbit, Jiggins, & Mallet 2001).

Many well-known examples of sexual selection and speciation are taken from birds. In some species, elaborate plumage color or morphology evolve by sexual selection and diversification of plumage characteristics is associated with speciation rate in some lineages (Price-Waldman, Shultz, & Burns, 2020; Beltran, Shultz, & Parra 2021). Studies characterizing the genetic basis of plumage differences in hybridizing birds shows that early stages of divergence are characterized by very high levels of genetic differentiation almost exclusively at plumage-associated regions, which tend to be clustered in few regions of the genome (Toews et al. 2016A; Poelstra et al. 2014). In finches, a rapid radiation is associated with repeated divergence at few regions of the genome - most notably in a region involved in the regulation of melanin-based plumage traits (Campagna et al. 2017). While the correlation between divergent secondary sexual characteristics and genomic regions with elevated differentiation or divergence provides some circumstantial evidence of a role for sexual signals in speciation with gene flow, it is not clear that these signals experience restricted gene flow or whether the correlation is due to trait evolution that drives speciation or occurs following speciation (Price-Waldman, Shultz, & Burns 2020; Campagna et al. 2017).

Subspecies of the yellow-rumped warbler, *Setophaga coronata coronata* and *S. c. auduboni* are are genetically differentiated and morphologically distinct with respect to several melanin and carotenoid-based plumage characteristics (Toews et al. 2016B;

Brelsford, Toews & Irwin 2017; Hubbard 1969). Differences are maintained despite a long history of hybridization following secondary contact (Mila, Smith, & Wayne 2007). Recent analysis using restriction-site associated DNA sequence data shows that the genetic basis of these traits involves several genomic regions but little involvement of the sex chromosomes, and that melanin and carotenoid-based traits differ in the number of genetic regions associated with the phenotype (Brelsford, Toews & Irwin 2017). A study of the genetic architecture of reproductive isolation between *coronata* and *auduboni* indicates that the Z chromosome includes a large fraction of genome regions experiencing restricted introgression, along with the larger autosomes, and that most of these regions colocalize with peaks of differentiation (see Chapter 2). By combining detailed information about the genetic basis of divergent plumage traits and reproductive isolation in yellow-rumped warblers, we can gain insight into the role of secondary sexual characteristics in the process of speciation.

In this study, we build upon previous work to elucidate the genetic basis of plumage differences in *Setophaga coronata coronata* and *S. c. auduboni* by conducting an expanded genome-wide association study using whole genome sequence data from many more individuals sampled from allopatric populations and the hybrid zone. Additionally, by comparing the genetic architecture of plumage to reproductive barriers, we are able to provide information regarding the involvement of conspicuous, divergent secondary sexual characteristics, and potentially pre-mating barriers, in reproductive isolation between hybridizing taxa. We predict that, if secondary sexual characteristics are important in the maintenance of reproductive isolation, the genomic regions

associated with these traits will exhibit restricted gene flow between *coronata* and *auduboni* populations.

Methods

Data Collection

We used the whole-genome sequence data produced for previous studies of the yellow-rumped warbler hybrid zone (see Chapter 1 for detailed methods). In summary, genetic sequence data were obtained from male yellow-rumped warblers captured in the field between 2005 and 2021 and sequence data obtained from the NCBI Sequence Read Archive. For our original samples obtained from wild-caught birds, we prepared whole-genome libraries using the SeqWell Plexwell 384 LP kit and sequenced these libraries at the UCSD Institute for Genomic Medicine. We aligned reads to a *Setophaga coronata coronata* reference assembly and used the SNP calling and filtering procedure outlined in Chapter 2.

Genome-wide Association Study (GWAS)

For the GWAS, we supplied genotypes for all SNPs that passed our quality filters and included 940 male *coronata*, *auduboni*, and hybrids for which phenotypic data were collected. Phenotypic data were obtained by scoring plumage characteristics from photographs of birds taken at the time of capture. We scored 8 traits (extent of yellow on throat, saturation of yellow on throat, extent of black on auricular, saturation of black on auricular, extent of white chin corner, eye spot, eye line and wing patch) on a scale from 0 (*coronata*-like) to 100 (*auduboni*-like) for all birds with photographs of sufficient quality for each trait, and scored one trait (tail pattern) on a scale from 2 to 6. This

scoring system is adapted from Pyle (1997) which uses a letter-based system for the relatively discrete variation in tail pattern on the tail feathers. Because the extent and saturation of pigmentation are highly correlated for both *throat* ($r^2 = 0.98$) and *auricular* ($r^2 = 0.97$), we chose to use composite characteristics *throat* and *auricular*, which are average scores of extent and saturation for each respective trait, instead of assessing saturation and extent independently. Photographic examples of these traits are provided in Figure 3.1 and the plumage trait reference photos are included in the supplemental Figures S3.2 – S3.5.

We used GEMMA (Zhou and Stephens 2012) to perform the GWAS, using the univariate model, and used age as a covariate only for wing patch and tail pattern, because trait values are correlated with age. Because GEMMA requires that the supplied genotype data do not contain missing values, we used *Beagle* v5.4 (Browning, Zhou, & Browning 2018) to generate an imputed VCF file. To ensure that we retained only sites that passed our initial quality filters, and to take advantage of the maximum available information for imputation, we imputed using the full, unfiltered SNP data, then applied a filter to retain only SNPs that passed the initial quality filters.

We chose a significance threshold for reported p-values of $1e^{-7}$. In order to obtain this threshold, we performed 200 iterations of the GWAS for each trait, permuting the phenotype values of all individuals, and then counting the number of SNPs in these permutations that are reported with p-values less than $1e^{-7}$. On average, we obtained less than 2 SNPs with p-values below $1e^{-7}$ per 200 permutations of the GWAS (Figure S3.1). By applying a significance threshold of $1e^{-7}$, we obtain a false discovery rate of 2 per the

number of significant associations for each trait. The number of significant associations for each trait are given in Table 3.1.

Figure 3.1: Photographs of yellow-rumped warblers. Image A shows a male *coronata* (left) and a male *auduboni* (right) captured at the same site at Pine Pass. Images B, C, and D, show a male *coronata*, a male hybrid, and a male *auduboni* with labels indicating the five traits with a large number of significantly associated SNPs identified in the GWAS.

Candidate genes

To locate genes that intersect SNPs with significant trait associations, we used Bedtools v2.31.0 *intersect* to locate genes that overlap our SNPs of interest with a 20 kbp buffer upstream and downstream of each SNP. This distance cutoff was applied as a middle-ground approach to identifying genes, given that the majority of SNPs are within 10 kbp from the gene they affect, and some SNPs as far as 100 kbp from the nearest gene may still have an affect (Brodie, Azaria & Ofran 2018). We obtained gene locations by mapping the annotations of the zebra finch (*Taeniopygia guttata*; Rhie et al. 2021) to the *S. c. coronata* reference genome using Liftoff (Shumate & Salzberg 2021). We selected candidate genes for carotenoid and melanin-based traits with the criteria that the gene was within 20 kbp of the most significant association for one or more traits.

Plumage and reproductive barriers

We fit clines to plumage traits using the quantitative trait model in HZAR v0.2-5 (Derryberry, Derryberry, Maley & Brumfield 2014) using all male birds for which each trait was measured. To compare phenotypic clines to SNP clines, we extracted cline parameters for plumage-associated SNPs for which cline data were available from Chapter 2. Cline analysis can provide information about the strength of a barrier to gene flow as the width of a cline is inversely proportional to the strength of selection against hybrids (Barton & Gale 1993). To determine whether clines for a plumage trait are involved in reproductive isolation, we compared the width of the phenotypic and genotypic clines to the cline fit to genome-wide ancestry. Additionally, we investigated candidate genes for the carotenoid-based and melanin-based plumage traits to determine

if these genes exhibit restricted introgression by comparing the trait-associated region to the results from a test for restricted introgression into *coronata*-like and *auduboni*-like hybrids (see Chapter 2). This comparison allows us to evaluate evidence for introgression of our candidate genes into both parent populations independently.

Results

Genome-wide association study and candidate genes

The traits we measured vary in the number of SNPs or regions with significant associations. Binning SNPs into intervals separated by more than 500 kbp between the next significantly associated SNP results in 63 regions significantly associated with throat color. *Wing* and *tail* patterns exhibit only weak associations, with only 2 and 30 significantly associated SNPs, respectively. *Spot* and *line* appear to be influenced by one and two genetic regions, and a narrow region of chromosome 20 exhibits the strongest genetic association with both traits. *Auricular* is similar, with SNPs within chromosomes 1a and 20 exhibiting the strongest association. Additional regions of the Z chromosome and chromosome 4 show weak but significant associations with *auricular*. The distribution of SNPs associated with *corner* shows a very similar pattern to *throat*, with significantly associated SNPs on many of the chromosomes and in regions of elevated differentiation. SNPs with significant associations tend to be found in regions with elevated F_{ST}. Results for each trait are summarized in Table 3.1 and Figure 3.2.

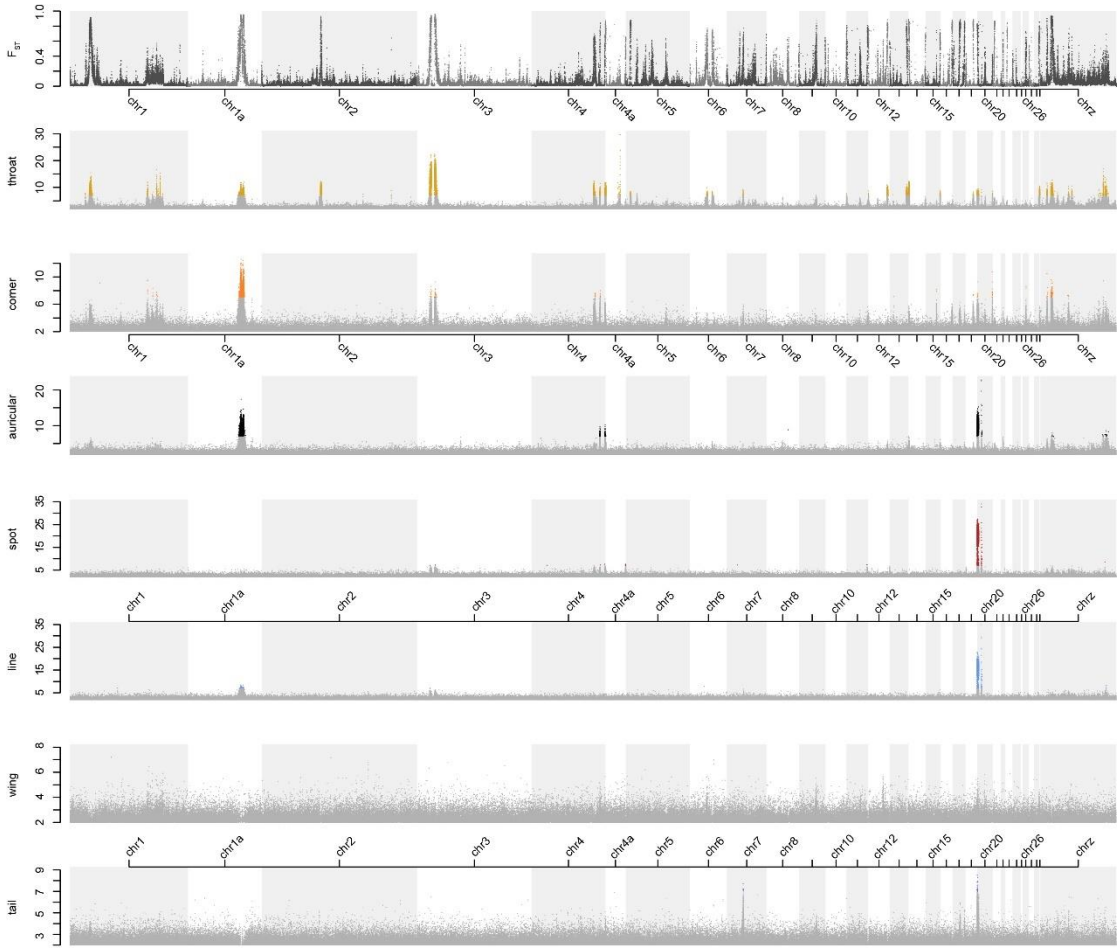

Figure 3.2: Differentiation (Weighted F_{ST}) between allopatric populations of *coronata* and *audubon*i measured in 10 kbp windows (top row), followed by -$\log_{10}$ transformed p-values for each of 7 traits. SNPs with p-values that pass the significance threshold of $1e^{-7}$ are colored.

To compare peaks of significant SNPs for different traits, we binned SNPs into intervals, where SNPs were assigned to the same interval if the p-value was above the significance threshold and the distance between them was at most 500 kbp. We find two regions that are particularly interesting due to the strength of the association with one or more traits and the narrowness of the region, relative to some broader peaks with strong trait associations. A region of chromosome 20 between approximately 4.19 and 4.37 Mbp stands out for its strong association with multiple traits as well as its narrowness. This region is associated with spot (p-value = 9.75 × 10-35), line (p-value = 1.18 × 10-35), auricular (p-value = 1.95 × 10-23), and throat (p-value = 5.32 × 10-10). This region of chromosome 20 intersects the gene ASIP, another well-known gene involved in melanic traits in birds (Campagna et al. 2017; Toews et al. 2016A; Uy et al. 2016) and in mammals (Kingsley, Manceau, Wiley, & Hoekstra 2009). These results are summarized in Figure 3.3A. On chromosome 4a, a region between approximately 14.36 and 14.46 Mbp contains a cluster of SNPs with significant association with throat color, which is white in *coronata* and yellow in *auduboni* (i.e. carotenoid-based pigmentation). The lowest p-value for SNPs in this region, and for any SNP associated with throat, is 2.15 × 10-30. This region intersects the gene EDA, a protein-coding gene known for its involvement in bony-plate development in sticklebacks and other derivatives of ectodermal tissues in mammals, such as hair and teeth (Colosimo et al. 2005; Thesleff & Mikkola 2002). This region has a significant association only with the throat color trait (Figure 3.3B). Despite strong association with plumage traits, these regions appear to experience a weak barrier to gene flow, shown by the weak restriction of introgression

(Figure 3.3, bottom half of each panel). To contrast the regions near ASIP and EDA, we provided a region of chromosome Z which experiences a strong barrier to gene flow, indicated by the highly significant signal of restricted introgression (Figure 3.3C) and no significant associations with any of the traits we measured.

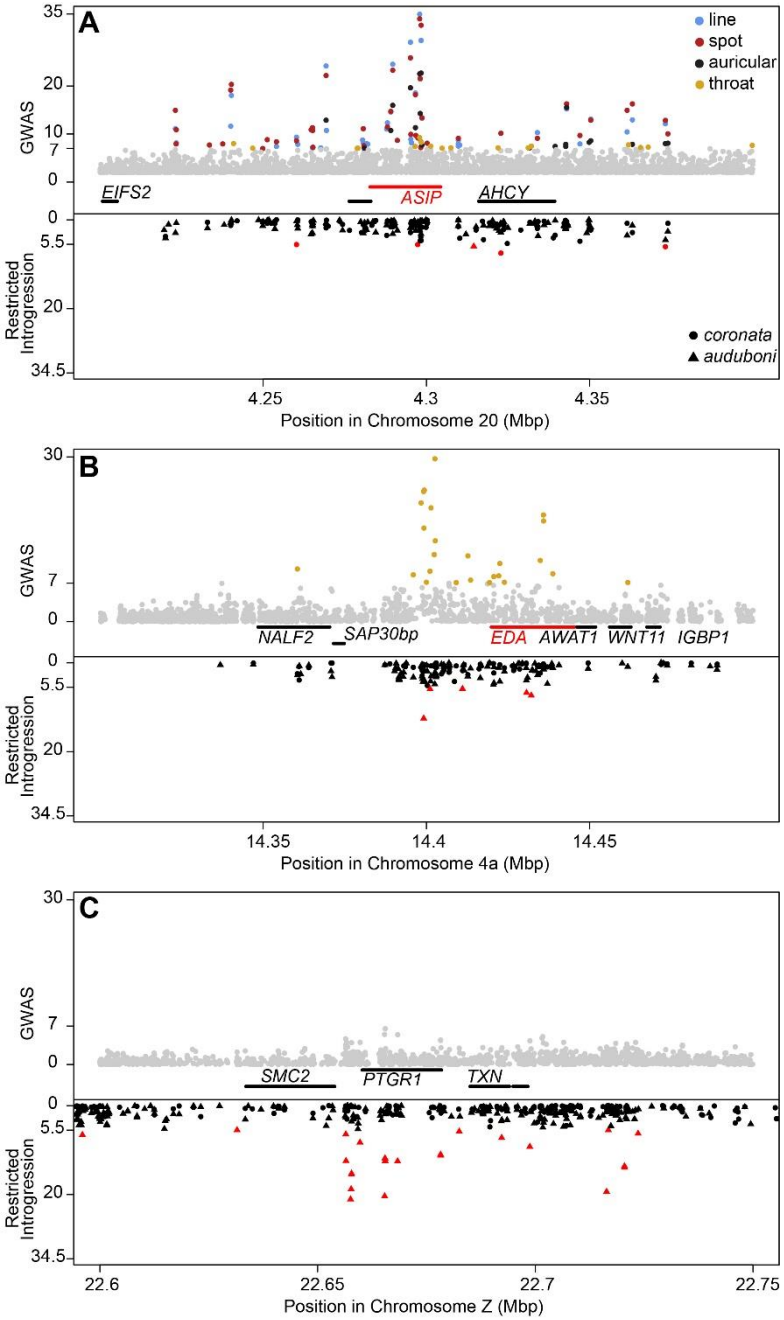

Figure 3.3: GWAS p-values for gene regions near SNPs with the strongest associations for spot, line, and auricular (A) and throat color (B). Below the GWAS results, p-values from the binomial test for asymmetric introgression is shown (see Chapter 2), where red points indicate SNPs with significantly restricted introgression (lower than expected frequency of heterospecific allele). In contrast, panel C shows a region with no significant trait associations but strongly restricted introgression into *auduboni*.

Phenotypic and genotypic clines for plumage traits

The genome-wide ancestry cline has a width of 170 km (158 km, 179 km; see Chapter 2). Parameter estimates for phenotypic and genotypic clines for all traits are summarized in Table 3.1 and Figure 3.4. In short, *throat, corner, line,* and *auricular* phenotypic clines are narrower than the genome-wide ancestry cline and clines of their significantly associated SNPs. Only *wing* and *tail* phenotype clines are wider than the ancestry cline. We did not obtain an estimate of cline width for SNPs associated with *wing* since none of the significantly associated SNPs were within genome regions that were included in the cline analysis. Cline center for traits and their associated SNPs are coincident and shifted eastward relative to the genome-wide ancestry cline with exception of *corner* and *spot*.

GWAS results				**Phenotype Cline (km)**		**SNP Clines (km)**	
Trait	**SNPs**	**Chromosomes**	**Intervals**	**Width (95% CI)**	**Center (95% CI)**	**Width (SD)**	**Center (SD)**
throat	15752	20	77	81 (70, 90)	6.4 (4.0, 9.3)	142 (29)	1.9 (4.9)
corner	1948	10	27	80 (50, 96)	-3.1 (-8.4, 2.9)	107 (35)	4.8 (2.5)
auricular	5047	6	20	68 (38, 98)	8.2 (2.4, 12.4)	122 (37)	3.9 (3.0)
spot	1583	7	12	114 (29, 220)	-6.9 (-14, 2.3)	157 (12)	2.1 (2.4)
line	1488	6	11	113 (24,155)	4.3 (-5.5, 11)	156 (15)	2.3 (2.6)
wing	2	2	2	178 (141, 241)	10 (-3.2, 19)	NA	NA
tail	30	2	2	224 (175, 278)	3.2 (-9.1, 12)	162 (15)	0.88 (2.3)

Table 3.1: Number of SNPs, chromosomes, and genomic intervals represented among significant associations for each trait, and cline parameters estimated for each plumage trait and their SNPs with significant associations.

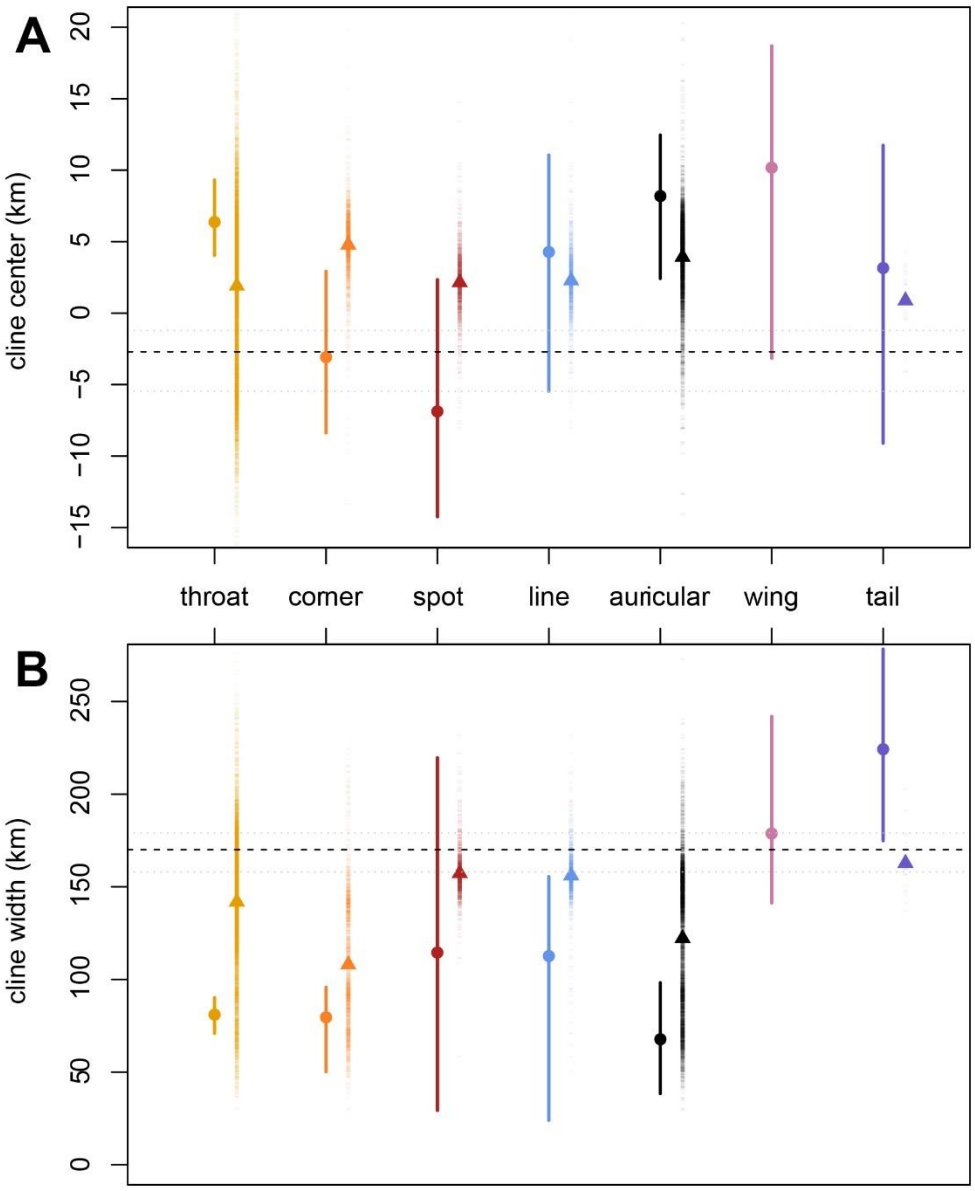

Figure 3.4: Summary of phenotypic (circles) and mean genetic cline parameters (triangles). 95% confidence interval for the phenotypic cline shown as a vertical line through the point estimate. Dashed lines represent the genome-wide ancestry cline, with the 95% confidence interval shown as a gray dotted line.

Discussion

The genetic architecture of plumage color

Similar to previous findings (Brelsford, Toews, & Irwin 2017), we find that the genetic architecture of plumage traits varies among the traits we measured. Throat color appears to be influenced by many genomic regions, while the white patches around the eyes (*spot* and *line*) share a common basis for their genetic control, indicated by highly significant associations with SNPs within the coding region of the gene *ASIP* on chromosome 20. We do find differences between the present study and the results of the 2017 study, however. The 2017 study identified a region on chromosome 15 as a candidate for control of carotenoid-based pigmentation. Our results show that this region, near the gene *SCARF2*, exhibits a relatively weak, but significant association with throat color (p-value = $9.7e^{-10}$). The strongest candidate gene involved in yellow pigmentation, based on the current results, is instead *EDA* on chromosome 4a. Significant associations are also found within relatively large regions of the genome, in some instances spanning several Mbp. For example, *spot* and *line* share an additional peak of significant SNP associations along the centromeric region of chromosome 20 from approximately 600 bp to 1.5 Mbp. Genetic control of throat color is more complex, with significant associations among 63 regions - many of which are quite broad. These peaks may include many genes and determining whether they are involved in plumage color will require more detailed analysis. Another recent study has identified other genes that are commonly associated with plumage coloration in *Setophaga* warblers, including the gene *BCO2*, which is

involved in carotenoid pigmentation (Baiz et al. 2021). Our GWAS does not associate this gene or any region of chromosome 24, where *BCO2* is located, with throat color.

The location of SNPs associated with *throat* and *corner* are highly correlated, which suggests a common molecular basis for these traits. This is somewhat unexpected, since for *throat*, the phenotypic difference between *coronata* and *auduboni* is the absence or presence of yellow pigmentation, while for *corner*, it is the absence or presence of gray pigmentation. The differences in the genetic architecture of these two traits may hint at differences in the genetic bases of carotenoid and melanin-based pigmentation. Wing and tail are melanin-based traits but we did not recover an association between *ASIP* and these traits, and find few significant associations in the genome overall. Genetic control of these traits is probably more dispersed among genes with very small effects.

Divergent plumage as a reproductive barrier

Gene flow is not substantially restricted for candidate genes with the strongest association with plumage traits, so it is unlikely that they have a strong role in reproductive isolation between *coronata* and *auduboni*. When comparing introgression of alternative alleles for *ASIP* and *EDA* regions, we find only marginally significant restricted introgression of alleles near and within *EDA* from *coronata* to *auduboni* and a similarly weak signal of restricted introgression of *ASIP* alleles in both directions (Figure 3.3). Clines for SNPs significantly associated with *spot* and *line* are marginally narrower than the genome-wide average cline (Figure 3.4B) and this pattern is driven mostly by a large region of chromosome 20, near the centromere. The narrow region containing SNPs with the strongest association with both spot and line has an average cline width of 177

km, 7 km wider than the genome-wide average of 170 km (however, this region includes few SNPs, and so is not reflected in the average of genetic clines shown in Figure 3.4A; see Chapter 2, Figures 2.3 & 2.6). The phenotypic clines for these two traits are narrower than their associated SNP clines. In other words, the phenotype for these traits transitions between parental types over a shorter distance within the hybrid zone than the frequency of alleles at strongly associated loci. This pattern suggests that if selection acts on plumage traits, selection on the trait itself is stronger than selection on any individual SNP that contributes to it.

If genetic control of a trait is non-additive, the phenotypic transition is likely to be abrupt and the center of the cline will be displaced toward the parental population in which the recessive allele is more common. The phenotypic cline center of *spot* is displaced to the west, relative to the genome-wide ancestry cline and SNPs associated with *spot*, while *line* shows the opposite pattern (Figure 3.4A, Table 3.1). In "pure" *coronata*, *spot* is apparent as a white patch anterior to the eye, and *line* is apparent as a white line extending from the eye toward the back of the head. These traits strongly contrast against the black background of the *auricular* in *coronata.* In *audubonì*, the *spot* and *line* regions of the face are uniformly gray. Other studies of the genes involved in melanogenesis in birds (Toews et al. 2016A) find that the melanic trait is recessive. The pattern observed in the phenotypic cline of *spot* is consistent with dominance of the *coronata* allele, since the displacement of the phenotypic cline is toward the west (within the *audubonì* side of the hybrid zone), relative to SNP clines and the genome-wide ancestry cline. However, the confidence intervals for the phenotypic clines of *spot* and

line are quite broad and, in the case of spot, include the estimates for their respective SNPs and the genome-wide ancestry cline, so it is more likely that discrepancies in the phenotypic and genotypic clines are due to measurement errors in scoring the phenotypes. Additional evidence that suggests that *spot* and *line* are additive traits: Brelsford, Toews & Irwin (2017) found genotype-phenotype associations are more consistent with additivity of *spot* and *line* in yellow-rumped warblers. Considering that *spot* and *line* associated SNPs do not experience a strong restriction of gene flow, as measured by our cline analysis and by comparing introgression within the genomic region with the strongest association with both traits, it is unlikely that *spot* and *line* are phenotypes involved in reproductive barriers.

There is, however, some support for a role for other plumage traits in reproductive barriers. Some plumage regions, such as the crown and throat, may be more important than others as sexual signals, (Price-Waldman, Shultz, & Burns, 2020). In yellow-rumped warblers, *throat* and *auricular* plumage regions are perhaps the most conspicuous differences and may be more important for signaling than subtle traits like *spot* and *line*. Phenotypic clines are consistently narrower than the genome-wide average cline, as well as the mean of clines fit to SNPs associated with the traits *throat, corner*, and *auricular*. These traits share some peaks of significant associations, with the most similarity among *throat* and *corner* (Figure 3.2). The similarities and differences in the genetic bases of these traits are reflected in the multimodal distributions of genotypic cline width estimates (Figures 3.3 & 3.4). The regions where significantly associated SNPs are found tend to be highly differentiated (Figure 3.2) but not all of these regions exhibit narrow

clines. Several peaks of differentiation on the large autosomes (chromosomes 1, 1a, and 3) and the sex chromosomes exhibit narrow clines and likely include genes that contribute to strong reproductive barriers (see Chapter 2). The large, highly differentiated regions of chromosome 1a and the Z chromosome show strong genotypic correlations, a very narrow cline, asymmetric introgression (*coronata* alleles are underrepresented in *auduboni*-like hybrids), and these regions are represented in the GWAS for *throat*, *corner*, and *auricular*, to varying extents. Plumage associated SNPs for *throat*, *corner*, and *auricular*, tend to cluster near peaks of differentiation, which are also near centromeres, and, with the exception of the peak for *auricular* on chromosome 20, are found within large "macrochromosomes" where recombination is relatively low (see Chapter 2). It is possible that plumage color is not an important factor maintaining reproductive isolation - instead, plumage genes may be hitchhiking due to physical linkage with other genes involved in reproductive barriers. On the other hand, physical linkage or linkage disequilibria among barrier loci can facilitate genome-wide divergence between lineages, and the coupling of sexual signals with ecological or intrinsic barriers can be an effective way to generate strong reproductive isolation.

The way in which plumage might act as a barrier to gene flow in yellow-rumped warblers is unknown. It is unlikely that it is involved in mate preference or choice, given that assortative mating within the hybrid zone is very weak or absent (Brelsford & Irwin 2009). Instead, plumage coloration may mediate aspects of male-male competition, in which birds with hybrid phenotypes are less effective at maintaining territories. Alternatively, male aggression biases toward homospecifics may promote divergence in

mating signals and facilitate sympatric coexistence (Seehausen & Schluter 2004; Dijkstra, Seehausen, Pierotti & Groothuis 2007). Whether male competition is important in maintaining divergence in yellow-rumped warblers will require detailed behavioral observations of the social interactions on their breeding territories.

The yellow-rumped warbler as a model of speciation with gene flow

The emerging picture of speciation in the *Setophaga coronata* complex is complex. Extensive sampling of both *coronata* and *auduboni* parental populations and their hybrids has enabled identification of the genetic basis of likely intrinsic post-zygotic reproductive barriers in high resolution and shown that reproductive barriers involve many loci distributed throughout the genome, that these barriers vary in the symmetry of their effects, and that they act together to maintain strong isolation (see Chapter 2). By contrasting this information against the genetic architecture of secondary sexual characteristics, we gain insight into the characteristics that may be or are not targets of selection. We find that genetic regions associated with sexual signals intersect regions that are involved in reproductive barriers, although the genes with the strongest correlation with plumage traits do not exhibit strong restriction of gene flow. Additionally, this hybrid zone offers the potential to investigate how dispersal and ecology influence hybrid zone dynamics. The long-term resilience of this hybrid zone, despite rapid ecological change and asymmetric gene flow, suggests that the hybrid zone is maintained by intrinsic reproductive barriers (see Chapters 1 and 2). However, transects through the hybrid zone exhibit some subtle and some distinct differences in cline shape that suggest topography, ecology, or demography are also important factors

that influence the permeability of the barrier to gene flow. The small, but growing collection of studies investigating the yellow-rumped warbler hybrid zone and the extensive data collection consolidated in this project provide a great deal of remaining potential to explore the evolutionary processes that generate biodiversity.

References

Baiz, M. D., Wood, A. W., Brelsford, A., Lovette, I. J., & Toews, D. P. (2021). Pigmentation genes show evidence of repeated divergence and multiple bouts of introgression in Setophaga warblers. Current Biology, 31(3), 643-649.

Barton, N. H., Gale, K. S., & Harrison, R. G. (1993). Genetic analysis of hybrid zones. Hybrid zones and the evolutionary process, 13-45.

Beltrán, D. F., Shultz, A. J., & Parra, J. L. (2021). Speciation rates are positively correlated with the rate of plumage color evolution in hummingbirds. Evolution, 75(7), 1665-1680.

Brelsford, A., & Irwin, D. E. (2009). Incipient speciation despite little assortative mating: the yellow-rumped warbler hybrid zone. *Evolution*, *63*(12), 3050-3060.

Brelsford, A., Toews, D. P., & Irwin, D. E. (2017). Admixture mapping in a hybrid zone reveals loci associated with avian feather coloration. Proceedings of the Royal Society B: Biological Sciences, 284(1866), 20171106.

Brodie, A., Azaria, J. R., & Ofran, Y. (2016). How far from the SNP may the causative genes be?. Nucleic acids research, 44(13), 6046-6054.

Browning, B. L., Zhou, Y., & Browning, S. R. (2018). A one-penny imputed genome from next-generation reference panels. The American Journal of Human Genetics, 103(3), 338-348.

Campagna, L., Repenning, M., Silveira, L. F., Fontana, C. S., Tubaro, P. L., & Lovette, I. J. (2017). Repeated divergent selection on pigmentation genes in a rapid finch radiation. Science advances, 3(5), e1602404.

Colosimo, P. F., Hosemann, K. E., Balabhadra, S., Villarreal Jr, G., Dickson, M., Grimwood, J., ... & Kingsley, D. M. (2005). Widespread parallel evolution in sticklebacks by repeated fixation of ectodysplasin alleles. science, 307(5717), 1928-1933.

Derryberry, E. P., Derryberry, G. E., Maley, J. M., & Brumfield, R. T. (2014). HZAR: hybrid zone analysis using an R software package. Molecular ecology resources, 14(3), 652-663.

Coyne, J. A., & H Allen Orr. (2004). Speciation. Sinauer Associates, Oxford University Press.

Dijkstra, P. D., Seehausen, O., Pierotti, M. E., & Groothuis, T. G. (2007). Male–male competition and speciation: aggression bias towards differently coloured rivals varies between stages of speciation in a Lake Victoria cichlid species complex. Journal of Evolutionary Biology, 20(2), 496-502.

Flaxman, S. M., Wacholder, A. C., Feder, J. L., & Nosil, P. (2014). Theoretical models of the influence of genomic architecture on the dynamics of speciation. Molecular ecology, 23(16), 4074-4088.

Hubbard, J. P. (1969). The relationships and evolution of the Dendroica coronata complex. The Auk, 86(3), 393-432.

Kingsley, E. P., Manceau, M., Wiley, C. D., & Hoekstra, H. E. (2009). Melanism in Peromyscus is caused by independent mutations in Agouti. PloS one, 4(7), e6435.

Maan, M. E., & Seehausen, O. (2011). Ecology, sexual selection and speciation. Ecology letters, 14(6), 591-602.

Mendelson, T. C., & Safran, R. J. (2021). Speciation by sexual selection: 20 years of progress. Trends in Ecology & Evolution, 36(12), 1153-1163.

Mila, B., Smith, T. B., & Wayne, R. K. (2007). Speciation and rapid phenotypic differentiation in the yellow-rumped warbler Dendroica coronata complex. Molecular Ecology, 16(1), 159-173.

Naisbit, R. E., Jiggins, C. D., & Mallet, J. (2001). Disruptive sexual selection against hybrids contributes to speciation between Heliconius cydno and Heliconius melpomene. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1478), 1849-1854.

Panhuis, T. M., Butlin, R., Zuk, M., & Tregenza, T. (2001). Sexual selection and speciation. Trends in ecology & evolution, 16(7), 364-371.

Poelstra, J. W., Vijay, N., Bossu, C. M., Lantz, H., Ryll, B., Müller, I., ... & Wolf, J. B. (2014). The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. Science, 344(6190), 1410-1414.

Price-Waldman, R. M., Shultz, A. J., & Burns, K. J. (2020). Speciation rates are correlated with changes in plumage color complexity in the largest family of songbirds. Evolution, 74(6), 1155-1169.

Pyle, P., Howell, S. N. G., Institute For Bird Populations, & Bird, R. (1997). Identification guide to North American birds : a compendium of information on identifying, ageing, and sexing “near-passerines” and passerines in the hand. Part I, Columbidae to Ploceidae. Slate Creek Press.

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., ... & Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), 737-746.

Ritchie, M. G. (2007). Sexual selection and speciation. Annu. Rev. Ecol. Evol. Syst., 38, 79-102.

Safran, R. J., Scordato, E. S., Symes, L. B., Rodríguez, R. L., & Mendelson, T. C. (2013). Contributions of natural and sexual selection to the evolution of premating reproductive isolation: a research agenda. Trends in ecology & evolution, 28(11), 643-650.

Seehausen, O., & Schluter, D. (2004). Male–male competition and nuptial–colour displacement as a diversifying force in Lake Victoria cichlid fishes. Proceedings of the Royal Society of London. Series B: Biological Sciences, 271(1546), 1345-1353.

Servedio, M. R., & Boughman, J. W. (2017). The role of sexual selection in local adaptation and speciation. Annual Review of Ecology, Evolution, and Systematics, 48, 85-109.

Shumate, A., & Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. Bioinformatics, 37(12), 1639-1643.

Toews, D. P., Brelsford, A., Grossen, C., Milá, B., & Irwin, D. E. (2016B). Genomic variation across the Yellow-rumped Warbler species complex. *The Auk: Ornithological Advances*, *133*(4), 698-717.

Toews, D. P., Taylor, S. A., Vallender, R., Brelsford, A., Butcher, B. G., Messer, P. W., & Lovette, I. J. (2016A). Plumage genes and little else distinguish the genomes of hybridizing warblers. Current Biology, 26(17), 2313-2318.

Thesleff, I., & Mikkola, M. L. (2002). Death receptor signaling giving life to ectodermal organs. Science's STKE, 2002(131), pe22-pe22.

Turelli, M., Barton, N. H., & Coyne, J. A. (2001). Theory and speciation. Trends in ecology & evolution, 16(7), 330-343.

Uy, J. A. C., Cooper, E. A., Cutie, S., Concannon, M. R., Poelstra, J. W., Moyle, R. G., & Filardi, C. E. (2016). Mutations in different pigmentation genes are associated with parallel melanism in island flycatchers. Proceedings of the Royal Society B: Biological Sciences, 283(1834), 20160731.

Xu, M., & Shaw, K. L. (2021). Extensive linkage and genetic coupling of song and preference loci underlying rapid speciation in Laupala crickets. Journal of Heredity, 112(2), 204-213.

Zhou, X., & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature genetics, 44(7), 821-824.

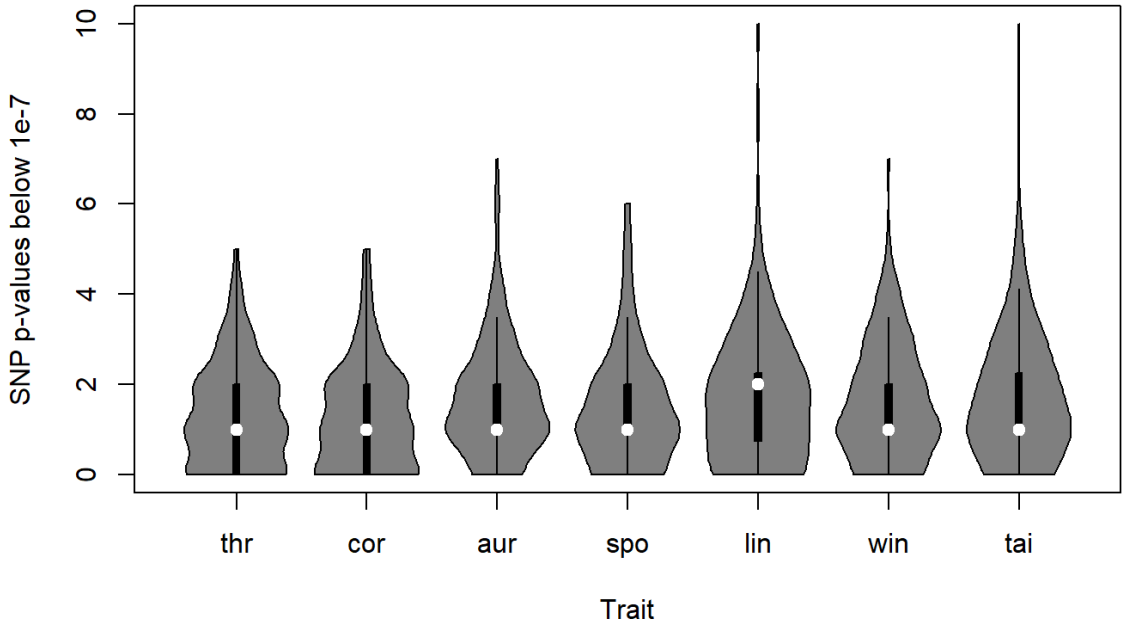

Figure S3.1: Results from 200 permutations of the GWAS. The number of significant associations, for a threshold of $1e^{-7}$, is shown on the y-axis. Kernel density is shown in gray.

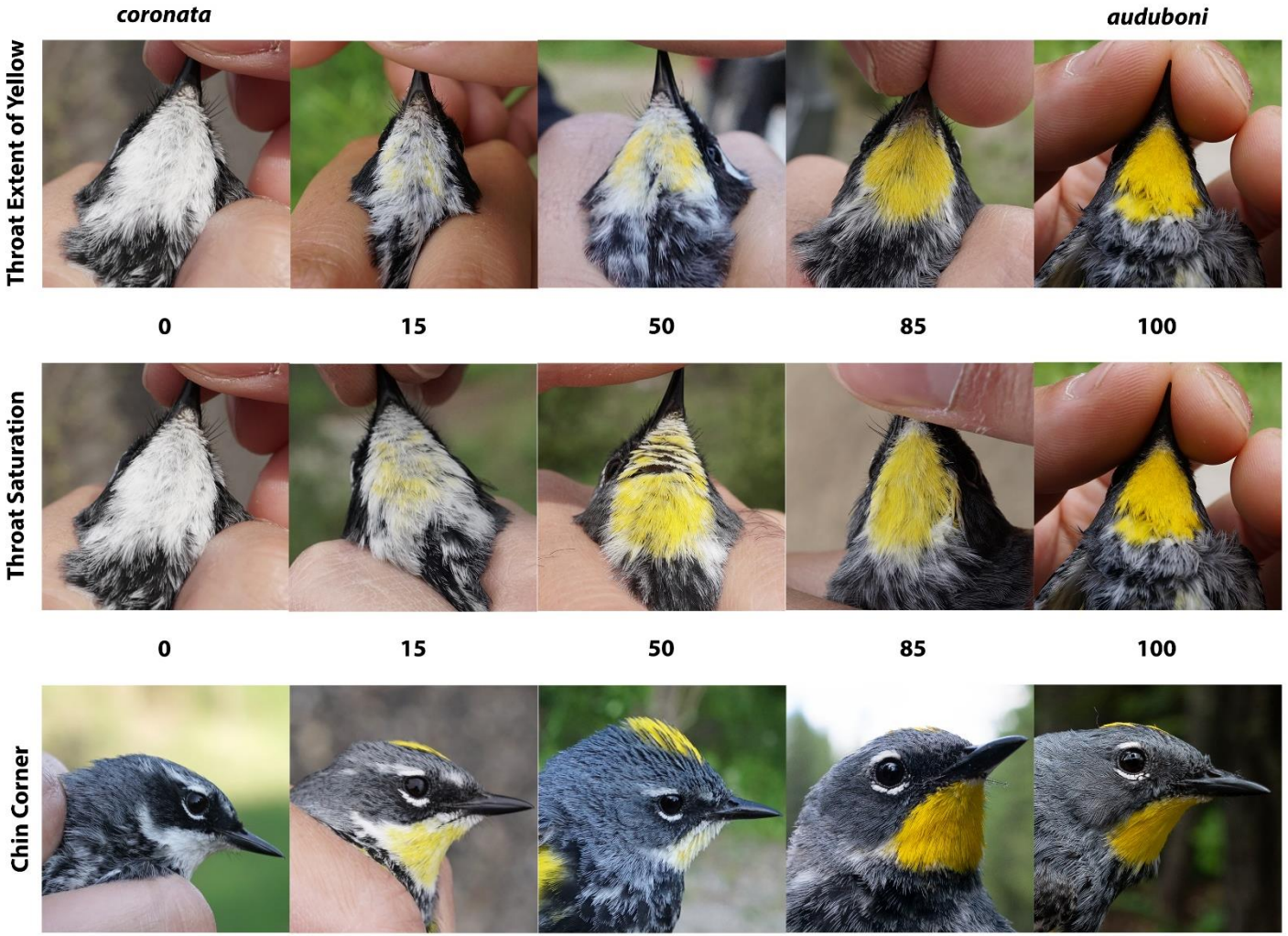

Figure S3.2: Plumage score references for throat color and chin corner.

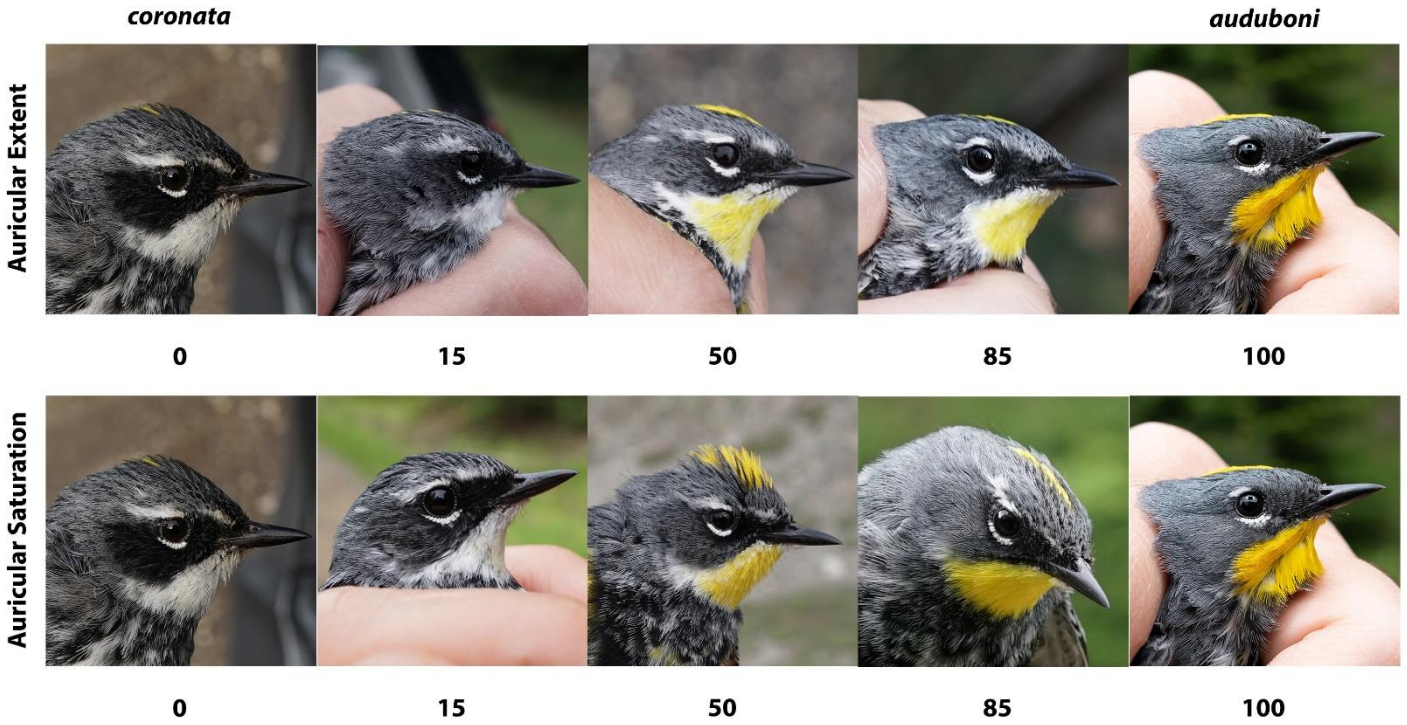

Figure S3.3: Plumage score references for auricular coloration.

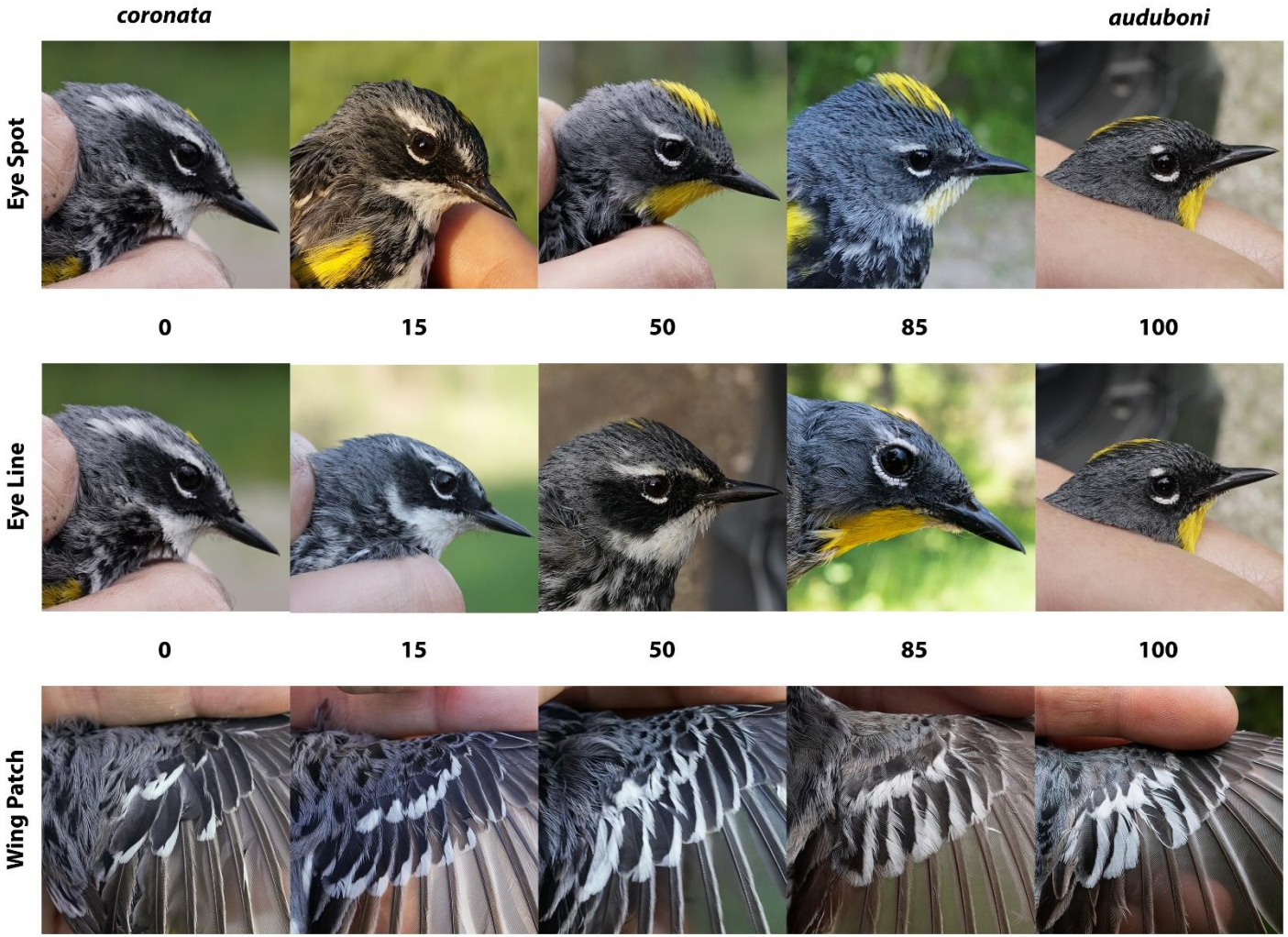

Figure S3.4: Plumage score references for spot, line, and wing patch.

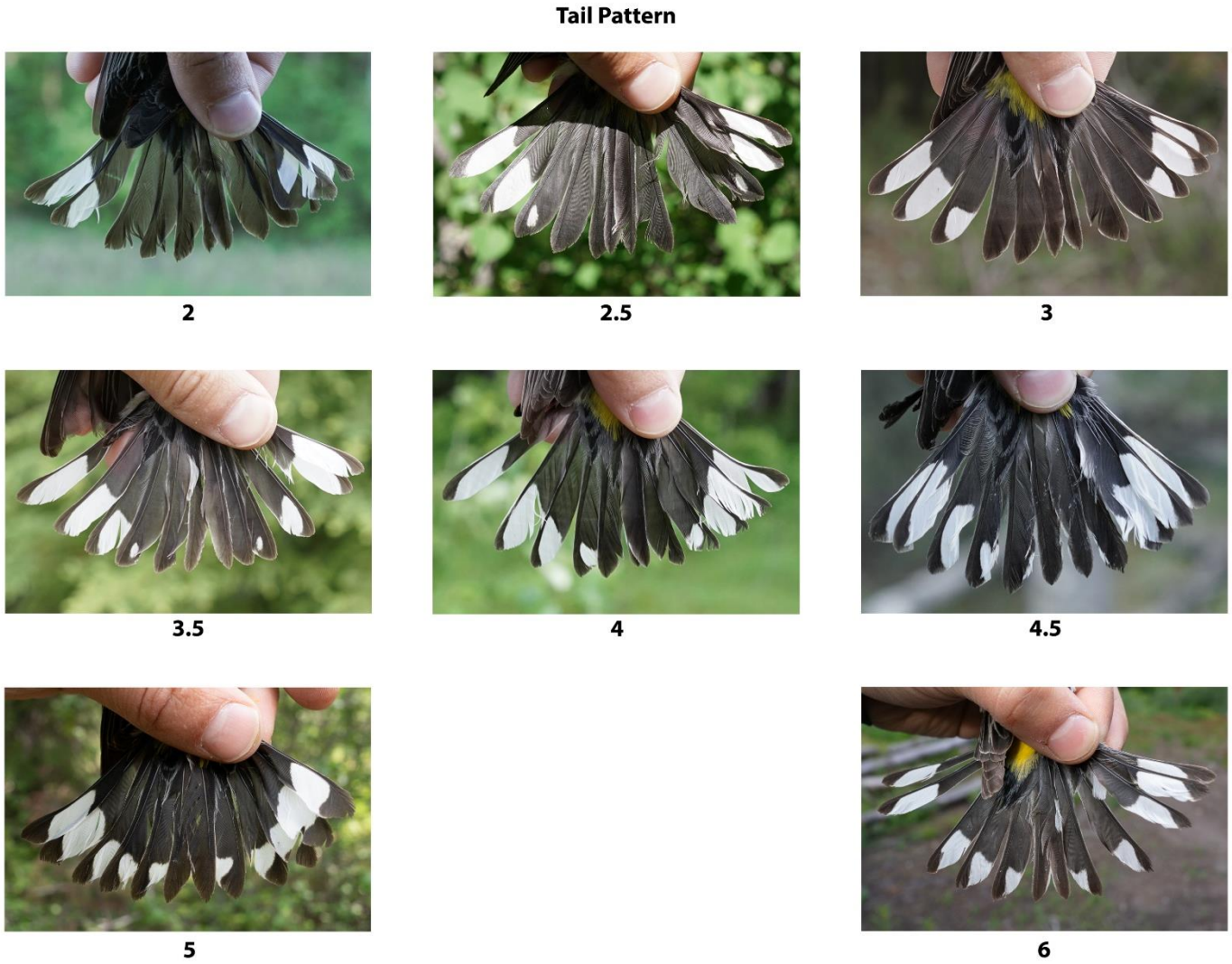

Figure S3.5: Plumage score references for tail pattern.

Figure S3.6: Male *S. c. auduboni*, EF06D02, captured near Mill Creek, Washington. Traces of *coronata*-like spot, line, auricular, and chin corner traits are apparent. Introgression of *coronata* ancestry can be found in birds far from the hybrid zone.