DISSERTATION

# ADVANCING CONSERVATION GENOMICS OF MIGRATORY SPECIES TOWARD A FULL ANNUAL CYCLE APPROACH

Submitted by

Matthew G. DeSaix

Department of Biology

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Fall 2023

Doctoral Committee:

Advisor: Kristen C. Ruegg

W. Chris Funk
David N. Koons
Peter P. Marra

ABSTRACT

ADVANCING CONSERVATION GENOMICS OF MIGRATORY SPECIES TOWARD A FULL ANNUAL CYCLE APPROACH

Global biodiversity loss is one of the foremost concerns of conservation efforts in the 21st century. The maintenance of genetic diversity within species is a critical factor in a species' persistence and adaptive potential in the face of changing environmental conditions. Migratory species make up more than 12% of the global vertebrate biodiversity and pose distinct challenges to conservation efforts due to inhabiting different geographical regions at different times of the year. The field of conservation genomics provides a valuable toolkit to addressing and understanding global biodiversity loss but requires additional methodological developments to better address the conservation challenges posed by migratory species. In my dissertation, I demonstrate advancements in conservation genomics aimed toward better understanding migratory species. In my first study, I addressed the question of ecological and genomic vulnerability to climate change in the Brown-capped Rosy-Finch (*Leucosticte australis*), an elevational migratory songbird of conservation concern. Second, I addressed a methodological gap in population genomics and developed statistical genetics models for using genotype likelihood data from low-coverage whole genome sequencing data to implement population assignment. In my last study, I demonstrate the utility of low-coverage whole genome sequencing for population assignment with detailing migratory connectivity in the American Redstart (*Setophaga ruticilla*). Altogether, my doctoral research demonstrates how genomic tools can help unravel the complexities of migratory species conservation. Furthermore, the species-

specific results are tied to knowledge gaps identified by wildlife managers and provide valuable information tied to conservation and management applications.

# ACKNOWLEDGEMENTS

I am immensely grateful to my advisor, Kristen Ruegg, who has been a wonderful mentor and advocate throughout my tenure as a PhD student and instrumental to my development as a conservation genomicist. Kristen's scientific curiosity and openness to a range of research questions inspired me to delve deeply into a variety of aspects of conservation genomics, while her skillful guidance has helped pull me out of the methodological weeds to understand the broader biological importance of my inquiries to conservation and the natural world. Furthermore, her commitment to fostering a lab environment that is open, welcoming, and fun stands out as one of the most enjoyable aspects of working with Kristen. I am fortunate to have been given the opportunity to be a part of this group and I cannot envision a more rewarding doctoral research experience than what I have received.

Additionally, I have received academic mentorship from a wide array of people. I am grateful to Eric Anderson for his mentorship, friendship, and enthusiastic support. Through Eric I have learned that rigorous scientific inquiry and outdoor recreation are not mutually exclusive endeavors, but rather complementary activities that produce excellent science when combined. Eric has played a pivotal role in building my foundation as a statistical geneticist and helping hone my craft in the technical aspects of bioinformatics, population genetics, and writing manuscripts out of a van.

I am grateful to my committee members, Chris Funk, David Koons, and Peter Marra, for providing critical review and guidance on my doctoral research. Specifically, David and Peter pushed me to extend my understanding of population genetic principles to broader biological aspects. I also owe Peter gratitude for funding my field work to Trinidad and Tobago, inviting

me to contribute to avian research outside my formal doctoral studies, and continually being a welcoming source of intellectual support. In addition to his role as a committee member, Chris has been an integral part of my broader doctoral research experience. I am grateful for his role in expanding my understanding of conservation genomics as well as his infectious fun-loving attitude and intellectual curiosity that makes working with him incredibly engaging.

At Colorado State University, I have been fortunate to be part of a wonderful scientific community that has profoundly impacted my research and personal well-being. Foremost, I am grateful to my incredible colleagues in the Ruegg Lab who have repeatedly and consistently been sources of support – Taylor Bobowski, Christen Bossu, Amanda Carpenter, Holden Fox, Jacob, Job, Caitlin Miller, Christine Rayne, Erica Robertson, Marina Rodriguez, Marius Somveille, Teia Schweizer, and Sheela Turbek. Every one of these people has contributed to making my doctoral research experience more fulfilling. Additionally, I am grateful for the friendship of my initial non-Ruegg Lab office mates, Julian Cassano and Nathan Phipps, and who are some of the most fun and compassionate people I know and helped me navigate my personal life in the initial years of graduate school. I am also grateful for Amir Alayoubi, Lisa Angeloni, Rebecca Cheek, Amanda Cicchino, Cole Deal, Lily Durkee, Brenna Forester, Bennett Hardy, Kim Hoke, Mary Linabury, Nico Matallana, Coby McDonald, Andrew Patton, Dan Sloan, Daryl Trumbo, Leena Vilonen, Miles Whedbee, and Mackenzie Woods for their influence on my research and life at CSU. Finally, I am grateful for the CSU Biology Department staff who have helped make all my research and conference travel possible.

I am also fortunate to have gotten to work with and learn from a wide range of collaborators outside of CSU. These people include, but are not limited to: Amy Seglund, Garth Spellman, Luke George, Erika Zavaleta, Ryan Harrigan, Caz Taylor, Jim Saracco, Nick Bayly,

Darshan Narang, Julie Hagelin, Lisle Gibbs, Thomas Sherry, Michael Webster, Thomas Smith, and Sean Hoban. I am particularly grateful to Amy Seglund and Julie Hagelin for providing funding resources and helping me understand the links between my research and the management needs of Colorado Parks and Wildlife and Alaska Department of Fish and Game, respectively. I am immensely grateful to Darshan Narang for facilitating and conducting field work in Trinidad and Tobago, as well as being a generous host for my stay there. In Trinidad and Tobago, I had the pleasure of working with and learning from Shivam Mahadeo, Richard Smith, Rachel Smith, and Carl Fitzjames. In addition, I am indebted and grateful to all field technicians and researchers who have contributed to our Brown-capped Rosy-Finch and American Redstart sample collection over the past 30 years.

My doctoral research was made possible by funding from Alaska Department of Fish and Game, Colorado State University Biology Department, Colorado Field Ornithologists, Colorado Parks and Wildlife, U.S. Department of Energy, The National Geographic Society, and The National Science Foundation. In addition, I am grateful to the generous donors to the Biology Department at Colorado State University from whom I received funding. I received additional scientific training from the University of Washington Summer Institute in Statistical Genetics and also Colorado State University School of Global Environmental Sustainability's Sustainability Leadership Fellows program.

Furthermore, numerous people provided me with the training and encouragement that led me to pursuing my dissertation work. Notably, Rodney Dyer, Lesley Bulluck, Catherine Viverette, and Andrew Eckert were amazing mentors during my M.S. degree at Virginia Commonwealth University and gave me a foundation in avian migratory research and population genetics. Thank you to Lindsey Miles for her patience in teaching me to write my first Bash

scripts and conduct bioinformatics analyses; Brandon Lind, Trevor Faske, Mitra Menon, Connie Bolte for teaching me evolutionary genetics: everything from cutting gels to critiquing papers; Bonnie Roderique and Jane Remfert for their camaraderie and support in the Dyer Lab; and Jessie Reese, Liz Schold, and Ben Nickley for being fellow bird-loving, fieldwork aficionados and helping me transition those skills to research.

Finally, I thank all of my friends and family outside of academia for their love and support, of whom I have too many to comprehensively list by name. I am honored to have such steadfast friends as Kenny Axford, Jesse Burgher, Phil Chaon, Charlie Coddington, Alex Darr, Jared Garland, John Moore, Kurt Ongman, and Francis Thomas who have supported me over the decades; Gabe Allen, Tim Fleeger, Nico Matallana, and Brian Orth for helping me keep a solid life-work balance in the Rocky Mountains; and to Marci Miller for training me to be disciplined with my mind and body. I am forever grateful to Eric Benson, Steven Barry, Tony Davis, Matt Gibson, Brian Kelly, and Mark Wagner for being personal mentors during my PhD studies and without whom this journey would not have been possible. I am also grateful to my family, especially my mother and father, Suzanne and John DeSaix, and sister, Kaila DeSaix, for always cheering me on and sending me their love in all I do. Thank you to my partner, Jessie Reese, for helping me embark on this journey and for providing unwavering support throughout. And to Faye, for all your love.

# DEDICATION

*To Jessie and Faye, for keeping the light burning*

TABLE OF CONTENTS

# INTRODUCTION

## Background

Global biodiversity loss is one of the foremost concerns of conservation efforts in the 21st century. Species' extinctions are a well-documented component of the Anthropocene era and are accelerating due to factors such as global climate change and habitat loss (Steffen et al., 2007; Ceballos et al., 2015; Urban 2015). Complementary to extinction, numerous extant species are declining in abundance at alarming rates and these widespread declines further threaten ecosystem functioning (Dirzo et al., 2014; Rosenberg et al., 2019). It is estimated that these population declines across taxa have resulted in a 6-10% loss of global genetic diversity (Leigh et al., 2019; Exposito-Alonso et al., 2022). While the maintenance of genetic diversity within species is a critical factor in a species' persistence and adaptive potential in the face of changing environmental conditions (Bernatchez 2016; Ceballos et al., 2017; Funk et al., 2019), this component has generally been lacking from global conservation initiatives (Laikre et al., 2020). However, as genomic tools continue to advance and become more widespread, the utility of genomics as an integral component of conservation and management practices is becoming increasingly apparent (Funk et al., 2019; Forester et al., 2022; Theissinger et al., 2023; Zamudio 2023).

Migratory species make up more than 12% of the global vertebrate biodiversity (Robinson et al., 2009) and pose distinct challenges to conservation efforts due to inhabiting different geographical regions or habitats at different times of the year (Runge et al., 2014). At each stage in the migratory annual cycle, migrant populations are subject to various stressors that can influence their fitness (Marra et al., 1998; Sillett et al., 2000). As a result, effective

conservation efforts require understanding migratory connectivity, defined as the links between different geographic regions used across the annual cycle (Marra et al., 2015; Webster et al., 2002). In the past 20 years, population genetics has become a well-established means for tracking migratory populations, especially for studies involving large sample sizes or small-bodied individuals (Faaborg et al., 2010). However, the value of genetic markers is often limited by the amount of genetic differentiation in a species and the availability of genetic data from individuals across the annual cycle (Faaborg et al., 2010; Lovette et al., 2004). Early methods relied on genetic markers that were limited to identifying only deep phylogeographic breaks within species (Kimura et al., 2002; Lovette et al., 2004; Ruegg & Smith, 2002). In recent years, next generation sequencing has facilitated the screening of a significantly larger number of genetic markers allowing for the delineation of breeding populations at finer spatial scales (Ruegg et al., 2014; DeSaix et al., 2023). Recent reductions in the cost of whole genome sequencing make genomics an increasingly cost-effective option for studying migratory connectivity and has the additional benefit of providing valuable data for assessing a species' adaptive potential.

Adaptive potential is a species' capacity to evolve in response to environmental change, thus species with greater adaptive potential have more resilience to stressors such as climate change. While quantifying adaptive potential has traditionally relied on experimental approaches to measuring the additive genetic variation underlying adaptive traits, population genomics vastly expands the ability to test for adaptation and identify the underlying genetic variation (Allendorf et al., 2010; Savoleinen et al., 2013). Importantly, while experimental studies are often restricted to model organisms that can be manipulated in laboratory setting, the use of population genomics to identify signals of local adaptation can be applied across taxa. For

example, genetic-environment association (GEA) methods are a common population genomics tool used to identify putatively adaptive loci in natural populations by searching for loci with allele frequencies that are associated with environmental variables (Rellstab et al., 2015; Forester et al., 2018). By identifying putatively adaptive variation and the environmental drivers of selection, researchers are able to further investigate the effects of global climate change on a species' adaptive potential (Razgour et al., 2019; Maier et al., 2023; Forester et al., 2023).

While assessing adaptive potential is already a difficult task for most species (Funk et al., 2019), this becomes further complicated in migratory species due to the dynamic nature of their populations' distributions throughout the year. A critical component of understanding how populations are locally adapted to their environment is accurately identifying the underlying environmental drivers that affect fitness. One solution to address this issue is using a priori knowledge of a species' life history to inform environmental variable selection in methods such as GEA (Hoban et al., 2016). With migratory species, identifying environmental drivers of adaptation requires knowledge of a species' life history across the annual cycle – including factors that affect fitness as well as the broader migratory connectivity patterns that determine where populations are present throughout the year. While comprehensive consideration of the adaptive potential of a migratory species has yet to be implemented, it is clearly an important step for the conservation of migratory taxa. Moving toward this objective requires additional research to improve existing genomic methods for investigating adaptive potential as well as developing new genomic tools for studying migratory species.

The overarching objective of my dissertation is to develop conservation genomics tools and methods to facilitate the study of migratory species. To that end, my dissertation includes different study systems and simulated datasets to best address my set of research questions. My

aim is to provide valuable results for the conservation of the species studied as well as address methodological aspects of the conservation genomics toolkit that are relevant across taxa.

**Research summary**

*Chapter 1*

In my first study, I addressed the question of ecological and genomic vulnerability to climate change in the Brown-capped Rosy-Finch (*Leucosticte australis*), a nomadic alpine songbird of conservation concern. Due to the alpine and nomadic life history of this species, the Brown-capped Rosy-Finch is one of the least studied landbirds in the United States. My work stemmed from initial research proposed by Colorado Parks and Wildlife, in collaboration with Colorado State University, University of California Santa Cruz, and Denver Museum of Nature and Science, to establish a baseline understanding of population abundance estimates and genetic health of this species. Our initial findings were published as a report to Colorado Parks and Wildlife (DeSaix & Ruegg, 2020) in which we used whole-genome sequencing data to detail that Brown-capped Rosy-Finches exhibit high gene flow across the breeding range with limited signatures of genetic isolation or inbreeding on any of the peripheral mountain ranges. Due to the concern of rapid climate change in alpine ecosystems, I directed my subsequent research with Brown-capped Rosy-Finch to investigate how this species may need to shift its distribution (ecological vulnerability) or adapt to changing conditions (genomic vulnerability) in response to climate change. I focused this research on a single stage of the annual cycle, the breeding range, to best document the methodological consolidations of this novel workflow before such work could be extended to examining the full annual cycle. This research was published in DeSaix et al. (2022) and our results highlighted that Brown-capped Rosy-Finch persistence may depend on

rapid adaptation to novel climate conditions in a contracted breeding range. Importantly, we also developed a metric for highlighting uncertainty in genomic vulnerability predictions due to novel climate conditions and demonstrated the need for studies to incorporate similar such metrics of uncertainty to highlight geographic regions for which predictions involving space-for-time substitutions are inappropriate (DeSaix et al., 2022).

*Chapter 2*

In my second study, I addressed a methodological gap in population genomics and worked with researchers at Colorado State University and the National Marine Fisheries Service to develop statistical genetics models for using genotype likelihood data from low-coverage whole genome sequencing data to implement population assignment. Population assignment methods are a standard tool for researchers to use genetic data to study migratory connectivity and no such methods had been developed to accommodate genotype likelihoods, or the computational burden of millions of genetic markers. Thus, population genomic study of the full annual cycle of migratory species was unable to take advantage of low-coverage whole-genome sequencing data until such methods were developed and made available in user-friendly software. To address this missing piece of the population genomics toolkit, I implemented our models in an open-source Python software package, WGSassign (https://github.com/mgdesaix/wgsassign), which my collaborators and I used to demonstrate highly accurate and computationally efficient population assignment with simulated and empirical data sets (DeSaix et al., *in review*). Specifically, we showed that WGSassign can provide highly accurate assignment, even for samples with low average read depths (< 0.01X) and among weakly differentiated populations. Furthermore, we derived the Fisher information

for allele frequency from genotype likelihood data and used that to describe a novel metric, the *effective sample size*, which figures heavily in assignment accuracy. Our development of WGSassign is an essential step for conservation genomics studies to be able to use genomic data from across the annual cycle in migratory species.

*Chapter 3*

In my final doctoral study, I used our development of WGSassign to investigate migratory connectivity in the American Redstart (*Setophaga ruticilla*). The American Redstart was an ideal study system for a comprehensive demonstration of the utility of low-coverage whole genome sequencing data for the study of migratory connectivity as there were previous migratory connectivity studies of this species to compare to. This study was an international collaboration that involved researchers from Colorado State University, National Marine Fisheries Service, SELVA Investigación para la conservación en el Neotropico, Trinidad and Tobago Field Naturalists' Club, State of Alaska Department of Fish and Game, The Ohio State University, The Institute for Bird Populations, Tulane University, Cornell lab of Ornithology, University of California Los Angeles Center for Tropical Research, and Georgetown University. We published this study in DeSaix et al. (2023) in which we revealed broad-scale parallel migration and highlighted unique population-specific patterns of connectivity in the American Redstart. By combining migratory connectivity results with demographic analysis of population abundance and trends, we provided full annual cycle conservation strategies for preserving numbers of individuals and genetic diversity (DeSaix et al., 2023). Notably, we highlighted the importance of the Northern Temperate-Greater Antilles migratory population as containing the largest proportion of individuals in the species. We further demonstrated the importance of

balancing the effective sample sizes of breeding populations to avoid assignment bias due to variation in the precision of allele frequency estimation. Overall, our results provide a valuable framework for studies that aim to use low-coverage whole genome sequencing data to understand the ecology and evolution of migratory species.

*Additional research*

In addition to the core studies enumerated in my dissertation, I have also collaborated on other projects pertaining to migratory species and conservation genetics. I worked with researchers at University of Maryland, Ithaca College, and Georgetown University on a mtDNA migratory connectivity study of the American Redstart which documented the nonbreeding range for two breeding populations that were delineated by phylogeographic patterns (DeSaix et al., 2022). I also collaborated with a researcher at Colorado State University in a study documenting a library preparation protocol for obtaining high quality whole genome sequencing data from feathers (Schweizer & DeSaix 2022). This protocol is a valuable resource for avian conservation genomics studies in which feather sample collection can be more feasible than blood collection, especially in large-scale international collaborations. Finally, I worked with researchers at Colorado State University, Alaska Department of Fish and Game, University of California Los Angeles Center for Tropical Research, The Institute for Bird Populations, University College London, and Tulane University on using migratory network models for conservation prioritization across the annual cycle. Migratory network models use migratory connectivity data (as can be obtained from WGSassign, Chapter 2) and population abundance data to quantify connectivity between stages of the annual cycle. My primary role on this project was developing

an R package, mignette, which makes these methods readily available for other migratory studies across taxa and using a variety of data sources (https://github.com/mgdesaix/mignette/).

**Conclusions and significance**

Addressing the biodiversity crisis for migratory species is a pressing concern in contemporary conservation efforts and requires multifaceted approaches. My doctoral research demonstrates how genomic tools can help unravel the complexities of migratory species conservation. My work with the Brown-capped Rosy-Finch (Chapter 1) resulted in methodological advancements for inferring genomic vulnerability to climate change and highlighting uncertainty in predictions to novel climate conditions. The results from this research provide a valuable step toward the application of genomic vulnerability methods across the full annual cycle of migratory species as they highlight the need to properly characterize the complexities of climate-based predictions. Subsequently, my work developing WGSassign for low-coverage whole genome sequencing data (Chapter 2) and comprehensively demonstrating its application to the study of migratory connectivity (Chapter 3) provide valuable contributions to the study of migratory species. For conservation and management projects that may have limited financial resources, low-coverage whole genome sequencing offers a potentially cost-effective approach for obtaining genomic data. The ongoing development of essential population genomic tools is an important step toward solidifying the role of genomics in conservation and management.

Furthermore, effective conservation-based research requires directly pairing research questions with the knowledge gaps identified by managers and policymakers (Knight et al. 2008; Hoban et al., 2013). In my doctoral research, the Brown-capped Rosy-Finch and American

Redstart studies provided valuable information to managers that was directly requested because they filled knowledge gaps in the management of these species. For the Brown-capped Rosy-Finch, the genetic results addressed Colorado Parks and Wildlife biologists' inquiries into the degree of gene flow and genetic diversity across the species' distribution. These results were an important component in determining that none of the sampled locations were from small, isolated populations that required immediate management action. The American Redstart migratory connectivity results informed Alaska Department of Fish and Game (ADFG) biologists that their breeding populations of American Redstarts likely overwinter in Mexico and Central America. The American Redstart is listed as a Species of Concern by the ADFG and the wintering information of breeding migratory birds in Alaska provides managers with the necessary information to target management resources. Collaborating with managers throughout my doctoral research was a vital component to having the research results linked to conservation and management application.

LITERATURE CITED


Allendorf, F. W., Hohenlohe, P. A., & Luikart, G. (2010). Genomics and the future of conservation genetics. Nature Reviews Genetics 2010 11:10, 11(10), 697–709. https://doi.org/10.1038/nrg2844

Bernatchez, L. (2016). On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. Journal of Fish Biology, 89(6), 2519–2556. https://doi.org/10.1111/JFB.13145

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human-induced species losses: Entering the sixth mass extinction. Science Advances, 1(5). https://doi.org/10.1126/SCIADV.1400253/SUPPL_FILE/1400253_SM.PDF

Ceballos, G., Ehrlich, P. R., & Dirzo, R. (2017). Biological annihilation via the ongoing sixth mass extinction signaled by vertebrate population losses and declines. Proceedings of the National Academy of Sciences of the United States of America, 114(30), E6089–E6096. https://doi.org/10.1073/PNAS.1704949114/

DeSaix, M. G., Connell, E. B., Cortes-Rodríguez, N., Omland, K. E., Marra, P. P., & Studds, C. E. (2022). Migratory connectivity in a Newfoundland population of the American Redstart (Setophaga ruticilla). The Wilson Journal of Ornithology, 134(3), 381–389. https://doi.org/10.1676/22-00004

DeSaix, M. G., George, T. L., Seglund, A. E., Spellman, G. M., Zavaleta, E. S., & Ruegg, K. C. (2022). Forecasting climate change response in an alpine specialist songbird reveals the importance of considering novel climate. Diversity and Distributions, 28(10), 2239–2254. https://doi.org/10.1111/DDI.13628

Dirzo, R., Young, H. S., Galetti, M., Ceballos, G., Isaac, N. J. B., & Collen, B. (2014). Defaunation in the Anthropocene. Science, 345(6195), 401–406. https://doi.org/10.1126/SCIENCE.1251817

Exposito-Alonso, M., Booker, T. R., Czech, L., Gillespie, L., Hateley, S., Kyriazis, C. C., Lang, P. L. M., Leventhal, L., Nogues-Bravo, D., Pagowski, V., Ruffley, M., Spence, J. P., Toro Arana, S. E., WeiÃŸ, C. L., & Zess, E. (2022). Genetic diversity loss in the Anthropocene. Science, 377(6613), 1431–1435. https://doi.org/10.1126/SCIENCE.ABN5642/

Faaborg, J., Holmes, R. T., Anders, A. D., Bildstein, K. L., Dugger, K. M., Gauthreaux, S. A., Heglund, P., Hobson, K. A., Jahn, A. E., Johnson, D. H., Latta, S. C., Levey, D. J., Marra, P. P., Merkord, C. L., Erica, N. O. L., Rothstein, S. I., Sherry, T. W., Scott Sillett, T., Thompson, F. R., & Warnock, N. (2010). Recent advances in understanding migration systems of New World land birds. Ecological Monographs, 80(1), 3–48. https://doi.org/10.1890/09-0395.1

Forester, B. R., Beever, E. A., Darst, C., Szymanski, J., & Funk, C. (2022). Linking evolutionary potential to extinction risk: applications and future directions. Frontiers in Ecology and Evolution, 20(9), 507–515. https://doi.org/10.1002/fee.2552

Forester, B. R., Day, C. C., Ruegg, K., & Landguth, E. L. (2023). Evolutionary potential mitigates extinction risk under climate change in the endangered southwestern willow flycatcher. Journal of Heredity, 114(4), 341–353. https://doi.org/10.1093/JHERED/ESAC067

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Molecular Ecology, 27(9), 2215–2233. https://doi.org/10.1111/MEC.14584

Funk, W. C., Forester, B. R., Converse, S. J., Darst, C., & Morey, S. (2019). Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists. Conservation Genetics, 20(1), 115–134. https://doi.org/10.1007/S10592-018-1096-1

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., & Whitlock, M. C. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. American Naturalist, 188(4), 379–397. https://doi.org/10.1086/688018/ASSET/IMAGES/LARGE/FG4.JPEG

Hoban, S. M., Hauffe, H. C., Pérez-Espona, S., Arntzen, J. W., Bertorelle, G., Bryja, J., Frith, K., Gaggiotti, O. E., Galbusera, P., Godoy, J. A., Hoelzel, A. R., Nichols, R. A., Primmer, C. R., Russo, I. R., Segelbacher, G., Siegismund, H. R., Sihvonen, M., Vernesi, C., Vilà, C., & Bruford, M. W. (2013). Bringing genetic diversity to the forefront of conservation policy and management. Conservation Genetics Resources, 5(2), 593–598. https://doi.org/10.1007/S12686-013-9859-Y/

Kimura, M., Clegg, S. M., Lovette, I. J., Holder, K. R., Girman, D. J., Milá, B., Wade, P., & Smith, T. B. (2002). Phylogeographical approaches to assessing demographic connectivity between breeding and overwintering regions in a Nearctic−Neotropical warbler (Wilsonia pusilla). Molecular Ecology, 11(9), 1605–1616. https://doi.org/10.1046/J.1365-294X.2002.01551.X

Knight, A. T., Cowling, R. M., Rouget, M., Balmford, A., Lombard, A. T., & Campbell, B. M. (2008). Knowing But Not Doing: Selecting Priority Conservation Areas and the Research-Implementation Gap. Conservation Biology, 22(3), 610–617. https://doi.org/10.1111/j.1523-1739.2008.00914.x

Laikre, L., Hoban, S., Bruford, M. W., Segelbacher, G., Allendorf, F. W., Gajardo, G., Rodríguez, A. G., Hedrick, P. W., Heuertz, M., Hohenlohe, P. A., Jaffé, R., Johannesson, K., Liggins, L., MacDonald, A. J., OrozcoterWengel, P., Reusch, T. B. H., Rodríguez-Correa, H., Russo, I.-R. M., Ryman, N., & Vernesi, C. (2020). Post-2020 goals overlook genetic diversity. Science, 367(6482), 1083–1085. https://doi.org/10.1126/SCIENCE.ABB2748

Leigh, D. M., Hendry, A. P., Vázquez-Domínguez, E., & Friesen, V. L. (2019). Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. Evolutionary Applications, 12, 1505–1512. https://doi.org/10.1111/eva.12810

Lovette, I. J., Clegg, S. M., & Smith, T. B. (2004). Limited Utility of mtDNA Markers for Determining Connectivity among Breeding and Overwintering Locations in Three Neotropical Migrant Birds. Conservation Biology, 18(1), 156–166. https://doi.org/10.1111/J.1523-1739.2004.00239.X

Maier, P. A., Vandergast, A. G., & Bohonak, A. J. (2023). Using landscape genomics to delineate future adaptive potential for climate change in the Yosemite toad (Anaxyrus canorus). Evolutionary Applications, 16(1), 74–97. https://doi.org/10.1111/EVA.13511

Marra, P. P., Cohen, E. B., Loss, S. R., Rutter, J. E., & Tonra, C. M. (2015). A call for full annual cycle research in animal ecology. Biology Letters, 11(8). https://doi.org/10.1098/RSBL.2015.0552

Marra, P. P., Hobson, K. A., & Holmes, R. T. (1998). Linking winter and summer events in a migratory bird by using stable- carbon isotopes. Science, 282(5395), 1884–1886.

https://doi.org/10.1126/SCIENCE.282.5395.1884/ASSET/EE1B5BFA-E0AB-4D1C-AAC2-66FB2EEE27A5/ASSETS/GRAPHIC/SE4887057004.JPEG
Razgour, O., Forester, B., Taggart, J. B., Bekaert, M., Juste, J., Ibáñez, C., Puechmaille, S. J., Novella-Fernandez, R., Alberdi, A., & Manel, S. (2019). Considering adaptive genetic variation in climate change vulnerability assessment reduces species range loss projections. Proceedings of the National Academy of Sciences of the United States of America, 116(21), 10418–10423. https://doi.org/10.1073/PNAS.1820663116/SUPPL_FILE/PNAS.1820663116.SD04.CSV
Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. Molecular Ecology, 24(17), 4348–4370. https://doi.org/10.1111/MEC.13322
Robinson, R. A., Crick, H. Q. P., Learmonth, J. A., Maclean, I. M. D., Thomas, C. D., Bairlein, F., Forchhammer, M. C., Francis, C. M., Gill, J. A., Godley, B. J., Harwood, J., Hays, G. C., Huntley, B., Hutson, A. M., Pierce, G. J., Rehfisch, M. M., Sims, D. W., Santos, M. B., Sparks, T. H., … Visser, M. E. (2009). Travelling through a warming world: climate change and migratory species. Endangered Species Research, 7, 87–99. https://doi.org/10.3354/esr00095
Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M., & Marra, P. P. (2019). Decline of the North American avifauna. Science, 366(6461), 120–124. https://doi.org/10.1126/SCIENCE.AAW1313/SUPPL_FILE/PAPV2.PDF
Ruegg, K. C., Anderson, E. C., Paxton, K. L., Apkenas, V., Lao, S., Siegel, R. B., Desante, D. F., Moore, F., & Smith, T. B. (2014). Mapping migration in a songbird using high-resolution genetic markers. Molecular Ecology, 23(23), 5726–5739. https://doi.org/10.1111/MEC.12977
Ruegg, K. C., & Smith, T. B. (2002). Not as the crow flies: a historical explanation for circuitous migration in Swainson's thrush (Catharus ustulatus). Proceedings of the Royal Society of London. Series B: Biological Sciences, 269(1498), 1375–1381. https://doi.org/10.1098/RSPB.2002.2032
Runge, C. A., Martin, T. G., Possingham, H. P., Willis, S. G., & Fuller, R. A. (2014). Conserving mobile species. Frontiers in Ecology and the Environment, 12(7), 395–402. https://doi.org/10.1890/130237
Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. Nature Reviews Genetics 2013 14:11, 14(11), 807–820. https://doi.org/10.1038/NRG3522
Schweizer, T. M., & DeSaix, M. G. (2023). Cost-effective library preparation for whole genome sequencing with feather DNA. Conservation Genetics Resources. https://doi.org/10.1007/S12686-023-01299-2
Sillett, T. S., Holmes, R. T., & Sherry, T. W. (2000). Impacts of a global climate cycle on population dynamics of a migratory songbird. Science, 288(5473), 2040–2043. https://doi.org/10.1126/SCIENCE.288.5473.2040/SUPPL_FILE/1049756.XHTML
Steffen, W., Crutzen, P. J., & McNeill, J. R. (2007). The Anthropocene. In Environment and Society (pp. 12–31). New York University Press. https://doi.org/10.18574/NYU/9781479844746.003.0006
Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P. R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G. R., Godoy, J. A., Jentoft, S., Malukiewicz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Secomandi, S., … Zammit, G.

(2023). How genomics can help biodiversity conservation. Trends in Genetics, 39(7), 545–559. https://doi.org/10.1016/J.TIG.2023.01.005/ATTACHMENT/EE319BA5-5B64-4402-9562-F4DA5AE168C2/MMC1.XLSX

Urban, M. C. (2015). Accelerating extinction risk from climate change. Science, 348(6234), 571–573. https://doi.org/10.1126/science.aaa4984

Webster, M. S., Marra, P. P., Haig, S. M., Bensch, S., & Holmes, R. T. (2002). Links between worlds: unraveling migratory connectivity. Trends in Ecology & Evolution, 17(2), 76–83. https://doi.org/10.1016/S0169-5347(01)02380-1

Zamudio, K. R. (2023). Conservation genomics: Current applications and future directions. Journal of Heredity, 114(4), 297–299. https://doi.org/10.1093/JHERED/ESAD019

# 1. FORECASTING CLIMATE CHANGE RESPONSE IN AN ALPINE SPECIALIST SONGBIRD REVEALS THE IMPORTANCE OF CONSIDERING NOVEL CLIMATE


## Summary

Species persistence in the face of climate change depends on both ecological and evolutionary factors. Here, we integrate ecological and whole-genome sequencing data to describe how populations of an alpine specialist, the Brown-capped Rosy-Finch (*Leucosticte australis*) may be impacted by climate change. We sampled 116 Brown-capped Rosy-Finches from 11 sampling locations across the breeding range. Using 429,442 genetic markers from whole-genome sequencing, we described population genetic structure and identified a subset of 436 genomic variants associated with environmental data. We modeled future climate change impacts on habitat suitability using ecological niche models (ENMs) and impacts on putative local adaptation using gradient forest models (a genetic-environment association analysis; GEA). We used the metric of niche margin index (NMI) to determine regions of forecasting uncertainty due to climate shifts to novel conditions. Population genetic structure was characterized by weak genetic differentiation, indicating potential ongoing gene flow among populations. Precipitation as snow had high importance for both habitat suitability and changes in genetic variation across the landscape. Comparing ENM and gradient forest models with future climate predicted suitable habitat contracting at high elevations and population allele frequencies across the breeding range needing to shift to keep pace with climate change. NMI revealed large portions of the breeding range shifting to novel climate conditions. Our study demonstrates that forecasting climate vulnerability from ecological and evolutionary factors reveals insights into population-level vulnerability to climate change that are obfuscated when either approach is considered independently. For the Brown-capped Rosy-Finch, our results suggest that persistence may

depend on rapid adaptation to novel climate conditions in a contracted breeding range. Importantly, we demonstrate the need to characterize novel climate conditions that influence uncertainty in forecasting methods.

## Introduction

Global climate change is dramatically affecting biodiversity and extinction rates are accelerating across taxonomic groups (Urban, 2015). Alpine organisms that already inhabit the upper elevational reaches can be at particular risk from climate change driving upslope range shifts due to reduced potential to shift their range (Freeman et al., 2018; Sekercioglu et al., 2008), however, this risk may be tempered in regions that provide an abundance of microclimates (Seastedt & Oldfather, 2021). If range shift is not feasible, a species' long-term persistence in the face of climate change will likely depend on evolutionary or behavioral adaptation (Aitken et al., 2008; Forester et al., 2018; Hoban et al., 2016; Hoffmann & Sgró, 2011). Advances in ecological genomics are elucidating the genomic architecture of local adaptation (Hämälä & Savolainen, 2019; Savolainen et al., 2013; Tigano & Friesen, 2016) and providing insight into population-level responses to climate change (Bay et al., 2018; Dauphin et al., 2021; Fitzpatrick et al., 2021; Fitzpatrick & Keller, 2015; Rellstab et al., 2016; Ruegg et al., 2018). While common garden experiments are widely recognized as the best method for identifying signals of local adaptation (Kawecki & Ebert, 2004; de Villemereuil et al., 2016), ecological genomic approaches provide an alternative in species where common garden approaches are infeasible due to constraints related to life history and conservation status (*i.e.* threatened or endangered status).

Ecological niche models (ENMs) are used to assess vulnerability to climate change by forecasting the distribution of climatic conditions that characterize an organism's current range

(Guisan & Thuiller, 2005; Pacifici et al., 2015). Given the variety of terminology used in the literature surrounding ENMs, we will follow the guidelines set out by Sillero (2011) that ENMs model an organism's ecological niche and the resulting output maps forecast habitat suitability. Genomic offset is a complementary approach to predicting climate vulnerability and provides a relative measure of the magnitude of evolutionary adaptation required for a population to track changing climate conditions (Capblancq et al., 2020; Fitzpatrick & Keller, 2015; Rellstab et al., 2021). Genomic offset is based on identifying genetic-environment associations putatively underlying local adaptation and predicting future adaptive genetic composition based on the current genetic-environment associations (e.g. with gradient forest models; Fitzpatrick & Keller, 2015). However, genomic offset predictions may ignore key ecological factors (e.g. habitat suitability) that would affect persistence, especially for organisms with ranges that are experiencing drastic environmental changes. While genomic offset has predominantly been assessed independently of ecological factors (Capblancq et al. 2020, Rellstab et al. 2021; but see Chu et al., 2021; Gougherty et al., 2021; Nielsen et al., 2021), vulnerability to climate change is a multifaceted problem that should be assessed with multiple methodologies and data sources (Dawson et al., 2011). Integrating genomic offset and ecological niche models would provide an understanding of the ecological factors shaping where populations could persist, and the evolutionary factors underlying the amount adaptation required to persist there.

The objective of our study was to combine methods for predicting population-level response to climate-driven disruptions to habitat suitability and genomic adaptation to improve forecasting of climate vulnerability. We addressed this objective using the Brown-capped Rosy-Finch (*Leucosticte australis*), an alpine-obligate species endemic to the Southern Rocky Mountains (Wyoming, Colorado, and New Mexico) and part of a broader species complex

notable for specializations to alpine sky islands and arctic tundra (Johnson et al., 2020). While climate change broadly results in species shifting distributions poleward and upward in elevation (Chen et al., 2011; Parmesan & Yohe, 2003), the Brown-capped Rosy-Finch has limited potential for poleward range shift given the isolation of the Southern Rocky Mountains from other high-elevation mountain ranges and the presence of congeneric species that already inhabit those mountain ranges. Furthermore, Brown-capped Rosy-Finches already occupy nesting cliffs at the highest elevations (above 3,350 m) of the Southern Rocky Mountains, which limits the possibility for major upslope range shifts, though they occupy lower elevations during the winter months (Johnson et al., 2020). Recent genetic studies have suggested that mountain ranges do not function as geographic barriers to dispersal for the North American Rosy-Finch complex (Black Rosy-Finch [*Leucosticte atrata*], Grey-crowned Rosy-Finch [*Leucosticte tephrocotis*], Brown-capped Rosy-Finch) given the level of ongoing gene flow among these species (Drovetski et al. 2009, Funk et al. 2021). Ongoing gene flow among Brown-capped Rosy-Finch populations may be an important component that mitigates genomic offset and prevents genetic isolation.

Here, we outline a process to assess climate vulnerability that considers evolutionary (e.g. genomic offset) and ecological factors (habitat suitability; Figure 1). We aim to answer the question: How can estimates of genomic offset and habitat suitability be combined to improve forecasts of climate vulnerability? Using genome-wide sequence data, we assessed population genetic structure and estimated levels of inbreeding and genetic diversity in order to describe spatial genetic variation and appropriately inform subsequent genetic-environment association (GEA) analyses (Forester et al. 2018, Funk et al. 2019). We performed environmental variable selection to identify a subset of uncorrelated predictors for use in the GEAs and ENMs. We

developed ecological niche models (ENMs) using the environmental predictor data and presence-absence data from the citizen-science database eBird (Sullivan et al. 2009). Additionally, we identified a subset of genomic variants associated with the environmental data and used these data to model allelic turnover across the landscape with gradient forest (Ellis et al., 2012; Fitzpatrick & Keller, 2015). Using ensembles of global climate models for two time periods, 2041-2070 and 2071-2100 (AdaptWest Project 2021), we then forecast climate vulnerability in relation to genomic offset and habitat suitability. We demonstrate a novel application of the niche margin index (Broennimann et al. 2021) to highlight uncertainty in genomic offset predictions due to novel climate conditions.

Specifically, this study aimed to 1) characterize the magnitude of genetic change required to track climate change and where populations could persist to minimize genomic offset; 2) predict climate-driven habitat suitability shifts into the future, and 3) compare the underlying climatic drivers of, and spatial vulnerability to, genomic offset and habitat suitability. The integration of these approaches will provide a better understanding of evolutionary and ecological factors underlying species response to climate change and improve our ability to forecast climate change impacts on biodiversity.

## Methods

### *Field sampling and sequencing*

We sequenced feather and blood samples from 116 individuals spanning 11 sites across the Brown-capped Rosy-Finch breeding distribution (Table 1.1). Samples were collected during the breeding season of 2017 and 2018. Individuals from the Lost Man Lake and Independence Lake sites were combined as a single sampling unit for subsequent analyses based on their proximity (< 1 km) and the low sample sizes (5 and 1 individuals, respectively). Engineer

Mountain and Horseshoe Basin sites were also in close proximity (< 5 km), but we retained them as separate sampling units due to the larger number of individuals per site (8 and 18 individuals, respectively).

We extracted DNA from blood samples using the standard protocol for Qiagen DNEasy Blood and Tissue Kits and we modified the protocol to maximize DNA yield from feathers. Whole genome sequencing libraries were prepared following modifications of Illumina's Nextra Library Preparation protocol. Pooled libraries were sequenced on HiSeq 4000 lanes at Novogene Corporation Inc. All sequence data were quality filtered (GATK: McKenna et al., 2010; BCFtools: Li, 2011; Samtools: Li et al., 2009) and aligned (Burrows-Wheeler Aligner software; Li & Durbin, 2009) to a high-quality Brown-capped Rosy-Finch reference genome that was created by Dovetail Genomics through 10x de novo assembly and HiRise Scaffolding. The reference genome was created from liver samples of the Brown-capped Rosy-Finch (Denver Museum of Nature and Science samples DMNS52416 and DMNS52417). The reference genome was annotated with the most recent zebra finch annotations available (NCBI GCA_008822105.2) using the program Liftoff (Shumate & Salzberg, 2021). For the input into all subsequent analyses, we extracted high-quality single-nucleotide polymorphisms (SNPs; Supporting information).

*Population genetic structure*

We performed several analyses to describe geographic patterns of genetic variation. The presence of closely related individuals can skew signatures of population structure so we used KING (Manichaikul et al., 2010) to identify and remove individuals with up to second-degree relationships (kinship > 0.0884). PCA provides an efficient non-model-based method for

assessing population structure in high-dimensionality data sets (Patterson et al., 2006). We implemented principal components analysis (PCA) using the R package SNPrelate (Zheng et al., 2012) in R version 3.6.2 (R Core Team, 2019). Additionally, we estimated individual ancestry coefficients with the snmf function in the R package LEA (Frichot et al., 2014; Frichot & François, 2015), and tested a range of clusters from K=1 to 6 with 100 iterations each. Finally, we tested for effects of isolation by distance (linearized FST versus log10 geographic distance) with a Mantel test in the R package adegenet (Jombart, 2008). Pairwise FST was calculated in VCFtools version 0.1.13 (Danecek et al., 2011). Pairwise FST provides an estimate of genetic divergence between populations where higher FST values indicate higher divergence. Genetic divergence can increase through genetic drift but is homogenized by gene flow between populations. Thus, any patterns of high FST between sites can be used to identify potential barriers to gene flow. The interaction between levels of gene flow and effective population size can result in different patterns of nucleotide diversity and inbreeding. We calculated nucleotide diversity across 25,000 base-pair windows and individual inbreeding coefficients using VCFtools (Danecek et al., 2011). We estimated contemporary effective population size using the LD method from NeEstimator (Do et al., 2014).

*Bioclimatic variables*

Snow is a major component of weather that shapes alpine communities. Snow cover can insulate soils from extreme cold air temperatures (Neuner, 2014) and also dictate the length of the growing season (Jonas et al., 2008; Keller et al., 2005). In some alpine plant species, reductions of snow cover can result in increased frost damage and decreased plant production (Abeli et al., 2012; Baptist et al., 2010; Inouye, 2000). The Brown-capped Rosy-Finch feeds on a

variety of seeds throughout the year and on insects during the breeding season (Johnson et al., 2020; Martin et al., 1961; Packard, 1968; Warren, 1916). Elevation is an important component of the Brown-capped Rosy-Finch breeding range in relation to the presence of high elevation nesting cliffs (Johnson et al., 2020). To encapsulate the range of bioclimatic factors that may influence Brown-capped Rosy-Finch alpine breeding habitat, we obtained 32 bioclimatic variables and elevation from the AdaptWest Project at a 1 km resolution (Wang et al. 2016, AdaptWest Project 2021). Variable selection involved removing correlated variables (Pearson correlation coefficient > 0.75) and using expert opinion to select the most likely biological relevant predictor from correlated sets. To best represent the current time period that corresponds to our sampled data, we obtained the bioclimatic variables as means across the time period of 1991-2020 and we obtained the dataset at an appropriate resolution for Brown-capped Rosy-Finch breeding movements (1 km).

*Identifying putative adaptive variants*

We used two genetic-environment association (GEA) approaches to identify a set candidate SNPs that are associated with environment. First, we implemented the multivariate approach of redundancy analysis (RDA) as it performs well for detecting weak, multilocus signatures of selection (Forester et al., 2018). We performed RDA using environmental and elevation data from individual sampling locations as the predictor variables and individual genotypes as the response variables. To account for isolation by distance, we created Moran Eigenvector Maps (MEMs) from the geographic locations of sampling data and conditioned the RDA model on the MEMs. All RDA analyses were conducted with the R package vegan (Oksanen et al., 2013) and step-wise model selection was performed using the ordistep function.

Multicollinearity in the model was checked with variance inflation factors (VIF) and predictors with a VIF greater than 10 were removed (Zuur et al., 2010). RDA component contribution was used to determine the number of components included for identifying candidate SNPs. Candidate SNPs underlying local adaptation were identified by having p-values outside a three standard deviation cutoff (two-tailed p-value = 0.0027).

Second, we used latent factor mixed models (LFMM) as a univariate regression model to identify candidate SNPs associated with each of the predictor variables (Frichot & François, 2015). We set the number of K latent factors based on the results from the individual ancestry coefficient results. For each model, we set the false discovery rate to 0.05 and calibrated the p-values by setting the genomic inflation factor to achieve a flat p-value distribution with a peak at 0 (François et al., 2016). LFMM analysis was conducted in R using the LEA package (Frichot & François, 2015). SNPs identified in both RDA and LFMM were used as the candidate SNP set putatively underlying local adaptation. We identified chromosomal position and gene information of the candidate SNPs using the Bedtools 'closest' function (Quinlan & Hall, 2010) with the annotated *Leucosticte australis* genome. We identified candidate genes by selecting SNPs within 10,000 bases from genes of known function and tested for gene ontology enrichment with the chicken (*Gallus gallus*) genome using the Gene Ontology resource (Ashburner et al., 2000; Carbon et al., 2021; Mi et al., 2019).

Importantly, GEA analyses rely on the assumption that current allele frequencies are at equilibrium with the environment (Capblancq et al., 2020; Lasky et al., 2018). However, populations may experience an adaptational lag associated with historical environmental conditions (Browne et al., 2019). To test the influence of this assumption, we created two

candidate SNP sets based on two temporal periods: one that temporally encompassed our sample period (1991 – 2020) and one based on potential adaptational lag (1961 – 1990).

*Geographic distribution of putative adaptive variation*

We used the gradient forest algorithm to describe the associations of spatial, environmental, and genetic variables (Ellis et al., 2012; Fitzpatrick & Keller, 2015). Gradient forest is a machine learning method developed to model ecological community turnover in relation to environmental gradients by creating separate random forest models for each species (Breiman, 2001; Ellis et al., 2012). Community turnover is then identified by aggregating environmental predictor importance for each species. This concept has been extended to landscape genetics by substituting allele frequencies at genetic loci for species and modeling adaptive genetic composition across the landscape (Fitzpatrick & Keller, 2015). The turnover functions in gradient forest allow for inference of the environmental predictors driving observed changes in allele frequency (Fitzpatrick & Keller, 2015). We fit gradient forest models to environmental and spatial data as predictors for the 9 sampling sites with at least 6 individuals using the package gradientForest (Smith & Ellis, 2013). We modeled adaptive genetic variation turnover on the landscape using the candidate SNP set as the response variable. Model tuning was performed on the parameters *mtry* (random subset of predictors used in random forest) and *ntree* (number of trees grown in each forest; Hastie et al., 2009). We evaluated model performance with prediction accuracy calculated from the out-of-bag samples (Ellis et al., 2012). We tested model performance of the candidate SNPs against a randomized model of candidate SNP allele frequencies and a SNP set that included putatively neutral loci (Supporting

information). Using the top gradient forest model, we interpolated genetic composition across the remaining 1 $km^2$ cells from the breeding range for which we did not sample genetic data.

*Habitat suitability under climate change*

We created ENMs using the ensemble modeling approach in the R package *biomod2* (Thuiller et al. 2016; Supporting information). Presence-absence data were obtained from the eBird Basic Dataset (Sullivan et al., 2009) using the R package *ebirdst* (Strimas-Mackey et al. 2021). We used the same uncorrelated set of environmental predictor variables as in the GEA analyses. Models were trained on random subsets of 80% of the data with 10 replications for 5 algorithms (regression based methods: generalized linear model (GLM; McCullagh and Nelder 2019), multiple adaptive regression splines (MARS; Friedman 1991), and machine learning methods: gradient boosting trees (GBM; Elith et al. 2008), maximum entropy (Maxent; Phillips et al. 2006), artificial neural networks (ANN; Lek and Guégan 1999)). Given the focus of our subsequent analyses on temporal forecasting, we aimed to use a set of algorithms with balanced biases and avoided models that tend to project extreme outcomes (Beaumont et al., 2016). Model performance was based on total area under the receiver operator and the relation of specificity and sensitivity (true skills statistic, TSS). Only the top performing algorithms were included in the final ensemble model. Binary rasters of suitable/unsuitable habitat were created based on a TSS threshold that maximized the sum of specificity and sensitivity since this has been shown to effectively represent presence (Jiménez-Valverde & Lobo, 2007).

Future distribution was modeled for two time periods (2041-2070 and 2071-2100) and for four different Shared Socioeconomic Pathways (SSPs). The SSPs vary in the possible climate change challenges global socioeconomic policy will produce (O'Neill et al., 2016): SSP126 (low

challenges), SSP245 (medium challenges), SSP370 (high challenges), and SSP585 (high challenges). Given that the SSP585 scenario most closely tracks the recent climate predictions from the Intergovernmental Panel on Climate Change report (IPCC, 2021), we used the SSP585 results for all figures in the main body of the article and provided details of the other scenarios in the Supporting information. For all possible combination of time period and SSPs (8 combinations), we obtained 1 km resolution bioclimatic data from 13 General Circulation Models provided by AdaptWest (Wang et al. 2016, AdaptWest Project 2021). We test for upward elevational shifts in habitat suitability between current and future projections using a two-sample t-test for the elevation values in the suitable habitat binary rasters.

*Genomic offset to climate change*

Genomic offset estimates the magnitude of evolutionary adaptation needed for a population to keep pace with climate change (Capblancq et al., 2020; Rellstab et al., 2021). When using gradient forest models, genomic offset is calculated by the Euclidean distance between current genetic composition with the predicted genetic composition based on future environment (Fitzpatrick & Keller, 2015). We calculated the mean genomic offset for each cell across the different SSP and time period combinations of future climate. In gradient forest models, environmental values outside the range of the provided trained values from sampling sites result in extrapolation of genetic composition. We used the default method of linear extrapolation from the non-linear turnover functions in the gradientForest package (Smith & Ellis, 2013; Supporting information).

*Quantifying uncertainty in genetic-environment associations*

The niche margin index (NMI) is a metric that characterizes the distance from niche margins with 0 representing the margin, 1 being the maximum value within the niche, and decreasing negative values representing distance outside the niche (Broennimann et al., 2021). We use this concept to quantify the niche margins of the observed environmental data from our sampling sites and then measure NMI for all raster cells in the genomic offset predictions of future climate. In our usage of NMI, negative values represent regions with novel future climate conditions in relation to the current observed genetic-environment associations (i.e. at the sampling sites). Positive NMI values represent regions with future climate conditions that are currently experienced on the breeding range. Thus, genomic offset predictions in regions with positive NMI are based on the space-for-time assumption in the gradient forest models (Capblancq et al., 2020), while genomic offset predictions in regions with negative NMI indicate higher model uncertainty due to extrapolation in the gradient forest models.

## Results

### *Population genetic structure*

Whole-genome sequencing produced genomic data with an average 6x depth of coverage and variant filtering resulted in 429,442 SNPs for subsequent genetic analyses. We removed 12 individuals from the data set due to relatedness. Visualizing PCA results revealed weak clustering of Pike's Peak individuals from other sampling sites. The weak PCA clustering of individuals suggests low genetic differentiation among the sites, which was also supported by low pairwise FST values ranging from 0 to 0.042 (mean FST = 0.012). The Mantel test did not identify associations between genetic and geographic distance ($r = -0.003$, p-value = 0.42), but visualization of these pairwise comparisons revealed the Pike's Peak population had elevated

genetic differentiation compared to other site comparisons. Individual ancestry coefficients had the lowest cross-entropy values for K=1 clusters (cross-entropy = 0.870). Results for K=2 had only slightly higher cross entropy (0.872) and revealed separation of Pike's Peak individuals, similar to PCA results. Nucleotide diversity was similar across sampling locations (pi mean = 0.00053, range = 0.00047-0.00056; Table 1.1). The per-individual F inbreeding statistic was also similar across sampling locations (F mean = 0.11, range = 0.02 – 0.25; Table 1). Effective population size for the five sampling locations that had sufficient sampling size ranged from 108 – 403 (Table 1.1).

*Identifying putatively adaptive variants*

The final uncorrelated environmental variable set consisted of mean temperature of the warmest month (MWMT), precipitation as snow (PAS), and summer heat moisture index (SHM; mean summer temperature divided by summer precipitation), as well as elevation. For RDA, we retained the first MEM (MEM1) spatial predictor for accounting for population structure as it was uncorrelated with the other predictor variables and explained 42.4% of the spatial variation. Model selection in the RDA retained all predictor variables. RDA outlier SNPs putatively associated with climate were identified by loadings on the first constrained axis. We identified 2,040 and 2,045 candidate SNPs from the 1961 – 1990 and 1991 – 2020 environmental predictor data sets, respectively. In LFMM, we used a lambda of 0.7 to achieve the optimal distribution of p-values for each of the four predictor tests. With K=2 latent factors, we identified 4,844 and 4,502 candidate SNPs from the 1961 – 1990 and 1991 – 2020 environmental predictor sets, respectively. Intersecting the RDA and LFMM data sets identified 501 and 436 candidate SNPs for the 1961 – 1990 and 1991 – 2020 environmental predictor sets, respectively. Gene ontology

enrichment analysis identified 12 genes associated with the biological process glutamatergic regulation of synaptic transmission (Gene Ontology ID: 0051966, p-value = 2.69e-6, false discovery rate = 3.67e-2) and 6 genes associated with regulation of small GTPase mediated signal transduction (Gene Ontology ID: 0051056, p-value = 2.73e-6, false discovery rate = 1.86e-2).

*Geographic distribution of putative adaptive variation and habitat suitability*

Our evaluation of tuning parameters in gradient forest models identified the out-of-bag testing accuracy to reach convergence with 100 trees (*ntree* = 100). Using all predictors in each tree (*mtry = 5*) achieved the highest proportion of variance explained across the predictors. The comparison of the two time period predictor sets, with the corresponding candidate SNPs, revealed similar relative predictor importance. Therefore, we continued all subsequent analyses with the 1991-2020 predictor set and candidate SNPs. With the candidate SNP set, raw predictor importance was ranked in descending order of precipitation as snow (PAS), mean temperature of the warmest month (MWMT), summer heat moisture index (SHM), elevation, and MEM-1 (Figure 1.2a). The order and magnitude of importance in the top predictor variables was not reflected in the randomized candidate SNP set or the reference SNP set that included neutral variation (Figure 1.2a). Turnover functions for the predictors revealed mostly step-wise patterns of allelic turnover, except for sharp turnover between precipitation as snow values of 500 – 600 mm (Figure 1.2b-f). Sampling sites were most strongly separated in genetic composition turnover driven by precipitation as snow (Figure 1.3a).

Filtering eBird data resulted in 192 presence points and 4,973 absence points in the ENMs. The Maxent and GLM algorithms were used for the ensemble ENM as they had the

strongest ability in discerning species presence with high mean true skills statistic across runs (Maxent: 0.86 +/- 0.03 standard deviation, GLM: 0.86 +/- 0.03) and area under the receiver operator curve (AUC; Maxent: 0.97 +/- 0.01, GLM: 0.95 +/- 0.02). Environmental variable importance was similar among the algorithms with MWMT, PAS, and elevation as the most important variables. Binary rasters were created using a habitat suitability threshold of 0.03 derived by maximizing the specificity and sensitivity of the model. The highest values for habitat suitability were produced for the highest elevation portions of the breeding range with lower habitat suitability in the northwestern portions of the Rocky Mountains (Figure 1.3b).

*Genomic offset and habitat suitability under climate change*

The magnitude of genomic offset was highly variable across the breeding range with some of the lowest values in the southwestern mountains (Figure 1.3c). Some of the eastern mountain ranges had the largest concentration of high genomic offset values (Figure 1.3c). While the magnitude of genomic offset increased with climate scenario and time period, the spatial patterns of the relatively low and high genomic offset remained the same. The ENM models revealed that future suitable habitat broadly became more fragmented in the 2041-2070 time period (Figure 1.3d). Future suitable habitat shifted upward in elevation from baseline habitat suitability projections by a mean of 178 m (3,367 m to 3,545 m) across all raster cells ($t = -74.6$, $df = 35378$, p-value < 2.2e-16, 95% CI: 173.3, 182.7). Less severe climate scenarios showed reduced range contraction, and range contraction increased when forecasted to the 2071-2100 time period.

*Quantifying uncertainty in genetic-environment associations*

For the baseline time period, the majority of the geographic region for which we interpolated genetic composition was within or close to the niche margins derived from our sampling sites (Figure 1.4a). For the 2041-2070 time period, a larger portion of the range shifted outside the niche margins, broadly indicating a shift to novel climate conditions. Comparing the environmental data among time periods showed in overall decrease in future precipitation as snow (Figure 1.4b) and increases in mean temperature of the warmest month (Figure 1.4c) and summer heat moisture index (Figure 1.4d). The largest shift to novel climate conditions occurred with the temperature of the warmest month (Figure 1.4c). Combining visualizations of genomic offset, habitat suitability, and NMI showed that the central portion of the breeding range had the most uncertain genomic offset predictions due to climate shifts (Figure 1.5).

## Discussion

In this study, we evaluate climate change consequences related to disruptions of climate conditions putatively underlying local adaptation and habitat suitability on the breeding range of an alpine specialist, the Brown-capped Rosy-Finch. Persistence of Brown-capped Rosy-Finch populations in the face of climate change may depend on rapid adaptation in a contracted region of suitable habitat. We broadly demonstrate genomic offset predictions by themselves can be problematic for inferring vulnerability to climate change when 1) changes in habitat suitability preclude a population from persisting in a region of forecasted low genomic offset and/or 2) when there are widespread regions forecasted to experience novel climate conditions.

*Comparing climate drivers of habitat suitability and local adaptation*

For the Brown-capped Rosy-Finch, precipitation as snow, mean temperature of the warmest month, and elevation were the strongest predictors of habitat suitability. Our results forecast that the lowest elevational limits of suitable habitat for Brown-capped Rosy-Finches will contract to higher elevations. Similar forecasts of suitable habitat loss at lower elevations have been made for another alpine-obligate species complex (avian genus *Lagopus;* Scridel et al. 2021). Importantly, the upward elevational shift of predicted high habitat suitability may not necessarily correspond to a similar scale of actualized range contraction. A key factor that may mitigate climate change risks for alpine species, especially in the Rocky Mountains, is the highly heterogeneous topography of the alpine landscape (Seastedt & Oldfather, 2021). Alpine microtopography can result in thermal refugia along short horizontal distances that mimic air temperature changes of hundreds of meters upslope (Scherrer & Körner, 2010). The American Pika (*Ochotona princeps*) is an example of a small alpine species that can behaviorally adapt to suboptimal thermal regimes by using different microhabitats (Millar et al., 2018; Rodhouse et al., 2017). While the thermal tolerance of the Brown-capped Rosy-Finch is unknown, behavioral adaptation to microhabitat use may be an important component of their climate change response. Given that Brown-capped Rosy-Finches nest in cliffs (Hendricks, 1977; Packard, 1968; Sclater, 1912), small changes in nesting site selection (e.g. cliff aspect) could provide dramatic differences in the microhabitat climate. Research into Rosy-Finch microhabitat usage and physiology would provide much needed additional information regarding predicted response to climate change.

The amount of precipitation as snow appears to have biological importance for both local adaptation and the realized niche for the Brown-capped Rosy-Finch (Figure 1.2b). In alpine plant

communities, snow cover can have a large effect on flower abundance and the evolution of adaptive traits to reduce frost damage (Inouye, 2000). In turn, this could affect Brown-capped Rosy-Finches foraging in the breeding season as they feed on available insects and seeds from a wide-range of plant species and families (Johnson et al., 2020; Packard, 1968). Our findings of candidate SNPs having enriched gene ontology categories – synaptic transmission (GO:0051966) and GTPase mediated signal transduction (GO:0051056) – provide an avenue for future research to understand environmental selective pressures. Of the 18 genes we identified in these gene ontology categories, 8 genes (*CDC42SE2*, *RASA3*, *ITGB1*, *SLIT2*, *RASGEF1A*, *GRIK2*, *GRM3*, *NRXN1*) are associated with cognitive function, 3 with high-altitude adaptation (*GRM5*, *NRXN1*, *HCN1*), and 2 with feather color and morphology (*KITLG*, *GRM8*). The cognitive-associated genes *SLIT2* and *GRM3* have been identified as being important to the foraging and food-caching habits of a montane bird, the Mountain Chickadee (*Poecile gambeli*; Branch et al., 2022). White-tailed Ptarmigan (*Lagopus leucura*) is another alpine specialist with low genetic differentiation and range-wide adaptive divergence potentially associated with diet (Fedy et al., 2008; Zimmerman et al., 2021). For Brown-capped Rosy-Finch, further elucidating the connections between gene functions and local adaptation (e.g. linking genotypes and phenotypes) is an important next step in understanding the effects of climate change.

*Geographic patterns of climate vulnerability*

Our integrative forecast of range shift and genomic offset in the Brown-capped Rosy-Finch shows that climate vulnerability from decreased habitat suitability and increased genomic offset do not necessarily align spatially. For example, some of the northern mountain ranges had low to medium values of genomic offset (Figure 1.3c) but were not forecasted to have suitable

habitat in the future (Figure 1.3d). However, some southwestern regions with the highest genomic offset (Figure 1.3c) also showed high vulnerability to loss of suitable habitat in the future (Figure 1.3d). Broadly, these results show that interpretation from genomic offset predictions alone leave out important considerations of climate vulnerability. Furthermore, these results underscore the importance of using multiple measures of vulnerability for informing conservation and management (Dawson et al., 2011; Rellstab et al., 2021). For organisms that inhabit regions experiencing large climate shifts, even the lowest genomic offset values may indicate relatively large allelic shifts required for a population to retain optimal genetic-environment associations.

Alpine climate conditions are changing dramatically in the Southern Rocky Mountains, especially in relation to snowpack (Pederson, Gray, Ault, et al., 2011; Pederson, Gray, Woodhouse, et al., 2011) and summer temperature increases (Pepin et al., 2022). In our study, NMI results suggest that the central and northwest portion of the Brown-capped Rosy-Finch breeding range are shifting to novel climate conditions (Figure 1.4a). Of the bioclimate variables most tied to habitat suitability and putative adaptive variation, the amount of precipitation as snow is decreasing across the breeding range in the future (Figure 1.4b) and the mean temperature of the warmest month is dramatically increasing (Figure 1.4c). However, our characterization of change in these specific bioclimate variables is based on the ecological niche model predictions of range from eBird citizen science data. Importantly, citizen science data for this organism may be more likely to be collected at lower elevations that are more accessible to observers than the higher elevation portions of the breeding range. This sampling bias could over- (or under-) estimate the current distribution of the breeding range, as well as the distributions of climate values across future time periods (Figure 1.4b-d). Nonetheless, our NMI

measures, which were solely based on climate distance from the climate niche defined by our sampling sites, reveal large climate shifts across high elevation portions of the range (Figure 1.4a).

*Considerations in forecasting genomic offset*

Recent reviews have highlighted a number of key assumptions and limitations that need to be addressed in the ongoing development of genomic offset methods for effective use in conservation (Capblancq et al., 2020; Rellstab et al., 2021). Genomic offset approaches assume that similar future conditions will result in similar genetic composition (space-for-time assumption). While this assumption may be problematic (e.g. multiple genetic architectures underlying an adaptive optimum), novel future conditions further increase the uncertainty of population response due to predicted genetic composition from unobserved environmental conditions. To address part of the uncertainty temporal extrapolation, we used the niche margin index highlight these regions of extrapolation to future climate conditions (Figure 1.4a). Given the reliance of gradient forest methods on temporal extrapolation from nonlinear turnover functions, potentially to novel conditions, we strongly recommend future studies to provide some measure of this uncertainty. Another similar assumption is that populations are at adaptive equilibrium with the temporal period during which genetic-environment associations are being tested. Long-lived species (e.g. trees) are particularly prone to violate this assumption given that populations may have been established centuries ago with different selection pressures (Rellstab et al., 2021). For the shorter-lived Brown-capped Rosy-Finch, we tested for the potential influence of adaptational lag by comparing environmental predictors between two baseline environmental periods. While our results suggested limited differences in genetic-environment

associations with these two periods, additional study into the role of effective population size and genetic drift in adaptive (non)equilibrium in this this system may be insightful (Láruson et al., 2022).

Furthermore, incorporating factors of evolutionary adaptation into genomic maladaptation forecasting methods could further refine these predictions. Large populations with gene flow and minimal genetic drift are expected to have higher adaptive potential than small, isolated populations (Funk et al. 2019). Our results show that Brown-capped Rosy-Finches have relatively high genetic connectivity and previous studies have showed that there is introgression within the Rosy-Finch complex (Drovetski et al., 2009; E. R. Funk et al., 2021). Gene flow can promote the rapid spread of beneficial alleles among populations and also maintain standing genetic variation for novel selection pressures (Bernatchez, 2016; Tigano & Friesen, 2016; Yeaman, 2015). Given that genomic offset does not account for gene flow, estimates of genomic offset may overestimate or underestimate future maladaptation (Exposito-Alonso et al., 2017). In the case of the Brown-capped Rosy-Finch, understanding the influence of gene flow on adaptation to a changing environment is an important next step for incorporating these results into management decisions.

*Conclusion*

Here we show that the Brown-capped Rosy-Finch faces climate threats across their breeding range from changing habitat suitability and disruptions of genetic-environment associations. Future persistence may depend on rapid adaptation to novel climate conditions in a contracted breeding range. Expanding future research to forecast climate threats across the fall and wintering range would facilitate an assessment of climate vulnerability across the full annual

cycle. We also note the importance of identifying the potential for behavioral adaptation to alpine microrefugia that may mitigate climate change threats. The results of this study highlight the importance of combining multiple methods to characterize climate vulnerability in a more nuanced manner than provided by any of the methods alone.

**Tables and figures**

**Table 1.1.** ID = four letter abbreviation for sample locations used in the manuscript. Latitude and longitude specify the coordinates for sampling site locations and sample size specifies the number of individuals captured at these locations. MWMT is the mean temperature of the warmest month (°C), PAS is the annual amount of precipitation as snow (mm), SHM is the summer heat moisture index (calculated by dividing MWMT by the mean summer precipitation), and the last column is the elevation of the sampling site (m). Pi = mean nucleotide diversity across 25000 base pair windows. F = individual inbreeding statistics. Ne = effective population size for locations with sufficient sample size for the linkage disequilibrium method of calculation. Two sampling sites (Engineer Mountain and Horseshoe Basin) from which individuals were combined as a unit for analyses due to close proximity of the sites, indicated by *.

| Location | ID | Lat | Lon | N | MWMT | PAS | SHM | Elevation | Pi | F | Ne |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Devil's Causeway | DECA | 40.03 | -107.16 | 15 | 11.2 | 954 | 32.9 | 3508.6 | 0.00053 | 0.10 | 403 |
| Emma Burr Mountain | EBMO | 38.75 | -106.41 | 3 | 10.5 | 413 | 39.9 | 3743.9 | 0.00047 | 0.24 | - |
| Engineer Mountain | ENMO | 37.96 | -107.57 | 8 | 9.7 | 783 | 29.4 | 3825.7 | 0.00053 | 0.09 | - |
| Horseshoe Basin | HOBA | 37.94 | -107.55 | 18 | 9.3 | 820 | 28.1 | 3891.4 | 0.00056 | 0.02 | 236 |
| Independence Lake | LMIN* | 39.14 | -106.56 | 1 | 9.5 | 603 | 33.4 | 3937.2 | 0.00053* | 0.13* | - |
| Lake Agnes | LAAG | 40.47 | -105.89 | 15 | 10.1 | 927 | 27.1 | 3586.2 | 0.00055 | 0.04 | - |
| Lost Man Lake | LMIN* | 39.14 | -106.57 | 5 | 9.5 | 603 | 33.4 | 3937.2 | 0.00053* | 0.13* | - |
| Mt. Maxwell | MOMA | 37.24 | -105.14 | 7 | 10.0 | 598 | 25.2 | 3825.3 | 0.00053 | 0.15 | - |
| Mt. Evans | MTEV | 39.58 | -105.64 | 11 | 8.2 | 543 | 22.4 | 4163.8 | 0.00051 | 0.18 | 108 |
| Pike’s Peak | PIPE | 38.83 | -105.04 | 21 | 8.8 | 407 | 17.0 | 4066.0 | 0.00052 | 0.02 | 217 |
| Snowy Range | SNRA | 41.36 | -106.30 | 12 | 11.5 | 943 | 35.1 | 3406.7 | 0.00052 | 0.11 | 140 |

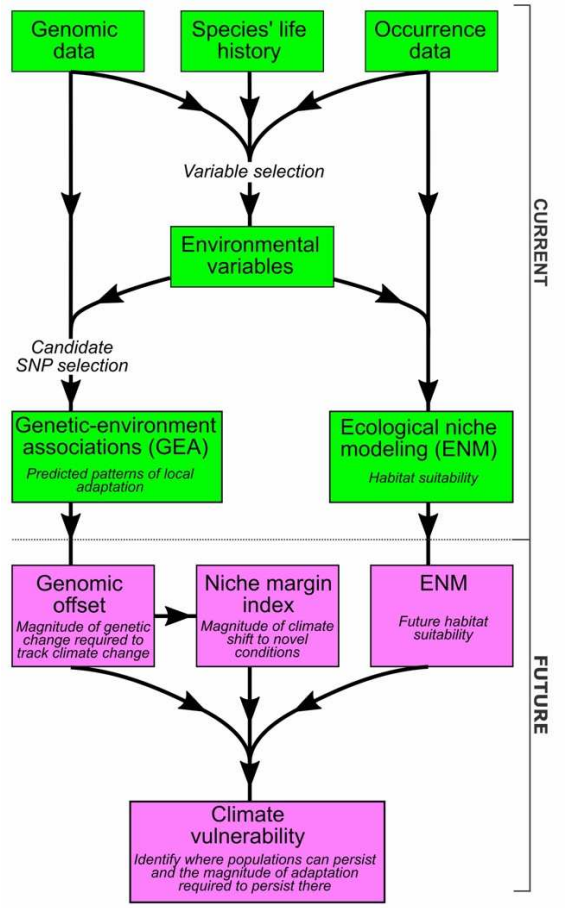


**Figure 1.1**. Workflow for combining ecological niche modelling and genomic offset for determining population-level climate vulnerability. Genomic data and occurrence data inform environmental variable selection by providing the geographic points for which environmental correlation is calculated. Species' life history information informs which variables are selected from correlated pairs, resulting in an uncorrelated set of environmental variables. Candidate adaptive SNPs are obtained through genetic- environment association outlier analyses using the environmental variables and genomic data. The resulting candidate SNPs are the input into gradient forest models which predict adaptive genetic composition across the landscape. The subset of environmental variables are also used with occurrence data in ecological niche models to habitat suitability. Gradient forest models are used to predict adaptive genetic composition to future climate and the distance with the baseline environment provides the measure of genomic offset. Additionally, the niche margin index is calculated to quantify the extrapolation to novel climate. Ecological niche models are also projected to future climate to provide a measure of future habitat suitability. The integration of these models provides a description of where populations are most likely to persist in the future and the magnitude of genetic change required to persist there. Furthermore, regions of novel climate are depicted to highlight uncertainty in the forecasting method.

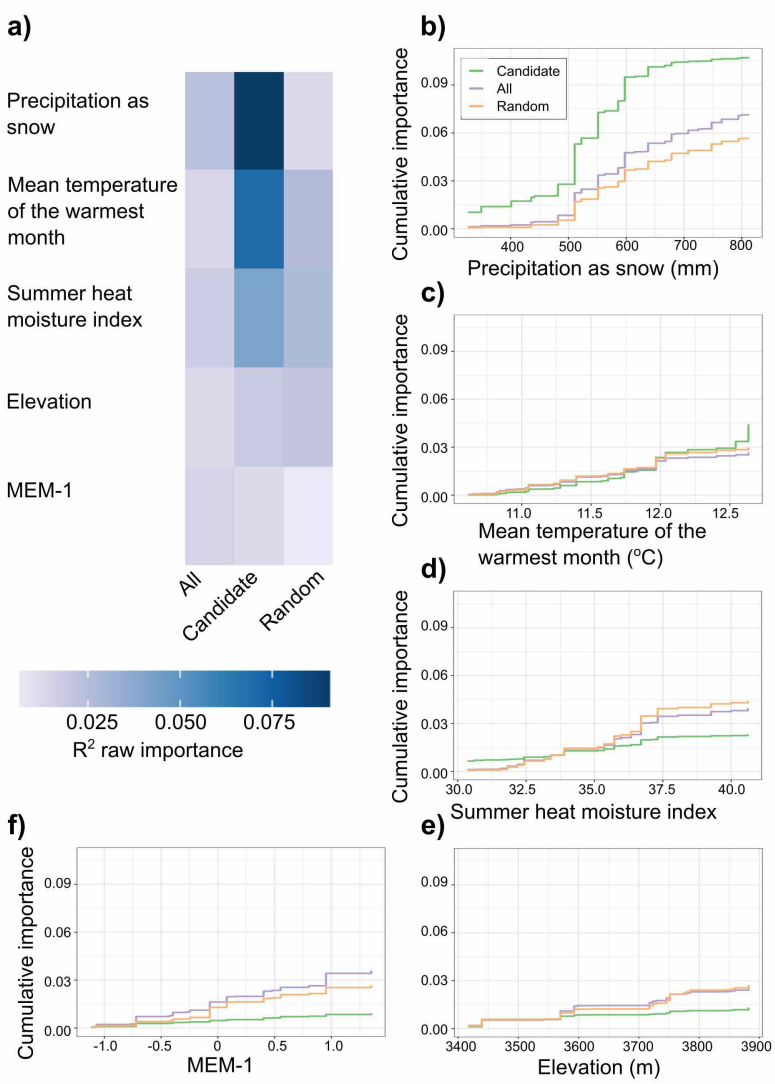


**Figure 1.2.** Performance of gradient forest models. (a) Raw R2 importance values for variables used as predictors in gradient forest model for three different datasets, which are: "All" is the total genomic variant set of 429,442 SNPs, "candidate" is the 436 candidate SNPs associated with the 1991–2020 baseline environment, and "random" is randomized genotypes of the candidate SNPs among the sampling locations. Using the candidate SNPs, larger raw importance values were obtained with the environmental predictors (precipitation as snow, mean temperature of the warmest month, and summer heat moisture index) than in the other two SNP sets. (b-f) the turnover functions from gradient forest model show the weighted cumulative importance values, which represent the relative importance of a variable in explaining changes in allele frequency. Here, only (b) precipitation as snow reveals consistently higher importance in the candidate SNP set than the other two datasets.

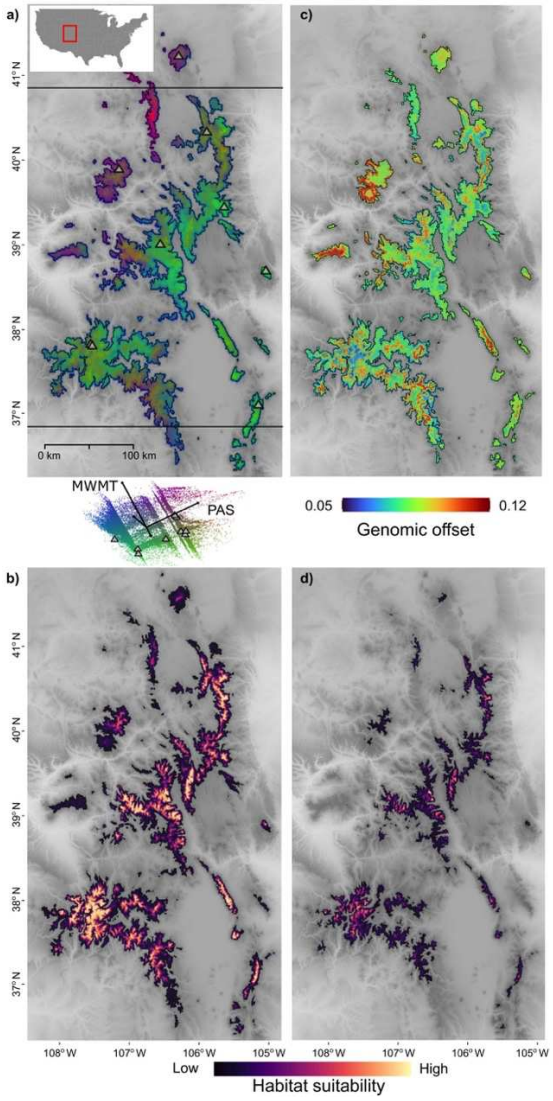


**Figure 1.3.** (a) Mapping of genetic composition from the candidate SNP gradient forest model with colours based on the biplot of environmental variable contribution to allele frequency change. Similar colours represent regions predicted to contain populations with similar genetic composition based on environment. Sampling sites represented by grey triangles. (b) Habitat suitability from the ENM for the current time period had the highest habitat suitability values in the highest elevation portions of the breeding range (bright yellow). The northwestern mountain ranges (e.g. snowy range and Devil's causeway) had some of the lower values of habitat suitability (darker colours). Using the future time period of 2041–2070 and the SSP 585 scenario we predicted genomic offset and habitat suitability. (c) Genomic offset was highly variable across the breeding range with some of the lowest values (blue) in the southwestern mountains and highest (red) in the eastern mountain ranges. (d) Habitat suitability decreased across the range with isolated patches of high suitability (bright yellow).

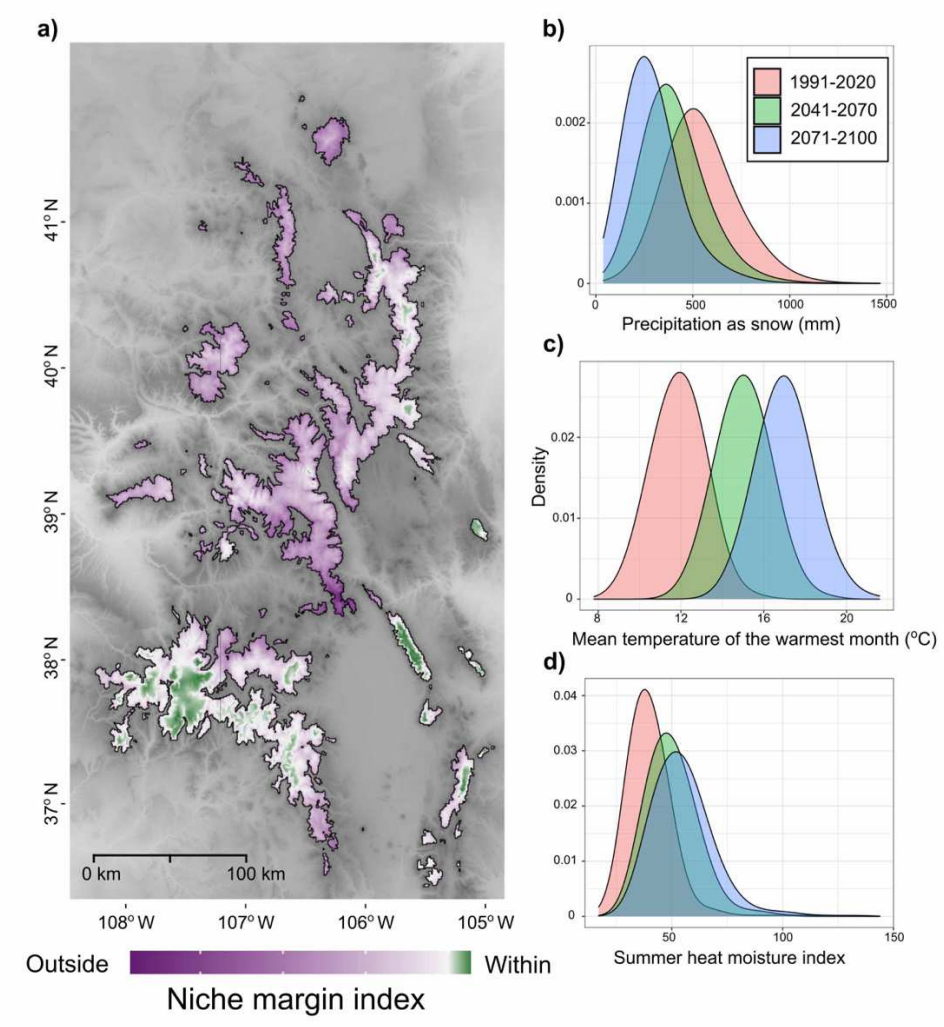


**Figure 1.4.** Identifying the magnitude of climate shift to novel conditions. (a) Calculating the niche margin based on our sampling sites and the niche margin index to future climate revealed large portions of the breeding range shifting to novel climate conditions (purple). The southern portions of the breeding range had the largest geographic areas retaining similar climate conditions (green) to the sampling sites. (b–d) of the three environmental variables in the gradient forest model that change temporally (i.e. excluding elevation), the largest shifts to novel conditions are present in the mean temperature of the warmest month.

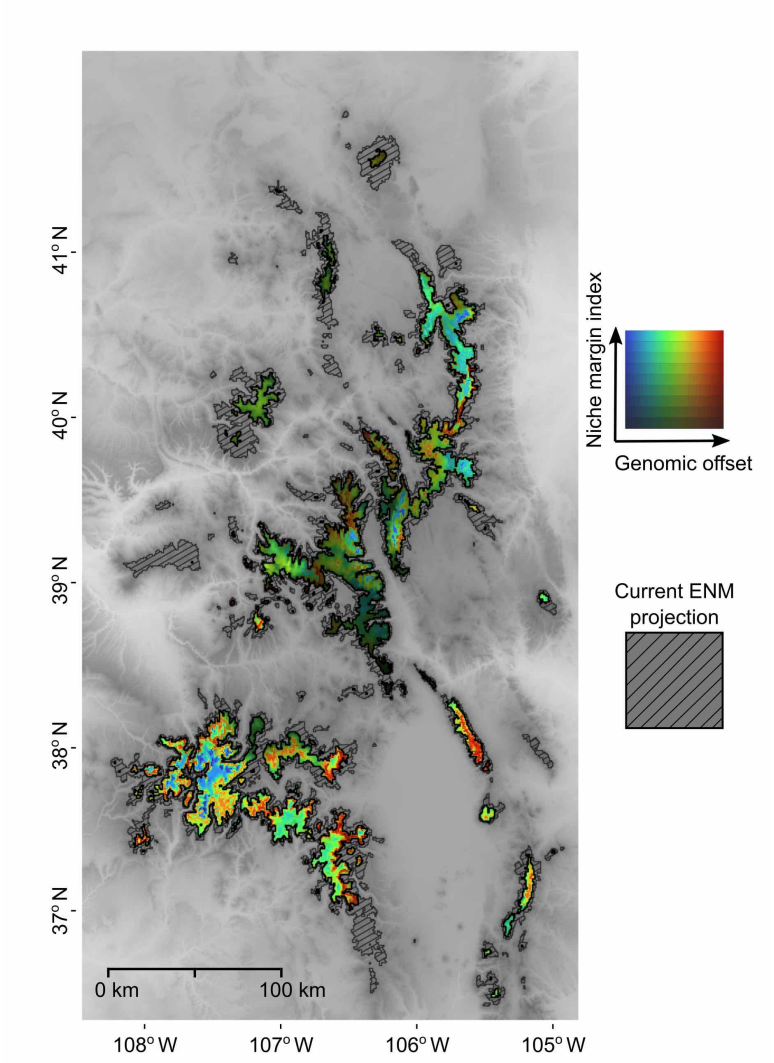


**Figure 1.5.** Population-level vulnerability to future climate of 2041– 2070. Colours represent genomic offset and the niche margin index (NMI). Genomic offset ranges from 0.05 (blue) to 0.12 (red), and the transparency of the colours reflects NMI. Bright colours represent NMI within the niche margins (between 0 and 1), while decreasing negative NMI values (novel climate) are represented by the darkening of the colours. Genomic offset predictions are shown for the predicted future suitable breeding range from the ecological niche model. The current ENM projection (1991– 2020) is shown with shaded black lines. The central and northwestern portions of the range have the largest concentration of regions shifting to novel climate, and therefore uncertain forecasting predictions.

LITERATURE CITED


Abeli, T., Rossi, G., Gentili, R., Mondoni, A., & Cristofanelli, P. (2012). Response of alpine plant flower production to temperature and snow cover fluctuation at the species range boundary. Plant Ecology, 213(1), 1–13. https://doi.org/10.1007/S11258-011-0001-5

Aitken, S. N., Yeaman, S., Holliday, J. A., Wang, T., & Curtis-McLane, S. (2008). Adaptation, migration or extirpation: climate change outcomes for tree populations. Evolutionary Applications, 1(1), 95–111. https://doi.org/10.1111/J.1752-4571.2007.00013.X

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics 2000 25:1, 25(1), 25–29. https://doi.org/10.1038/75556

Baptist, F., Flahaut, C., Streb, P., & Choler, P. (2010). No increase in alpine snowbed productivity in response to experimental lengthening of the growing season. Plant Biology, 12(5), 755–764. https://doi.org/10.1111/J.1438-8677.2009.00286.X

Bay, R. A., Harrigan, R. J., Buermann, W., Le Underwood, V., Gibbs, H. L., Smith, T. B., & Ruegg, K. (2018). Genomic signals of selection predict climate-driven population declines in a migratory bird. Science, 359(6401), 83–86. https://doi.org/10.1126/science.aat7956

Beaumont, L. J., Graham, E., Duursma, D. E., Wilson, P. D., Cabrelli, A., Baumgartner, J. B., Hallgren, W., Esperón-Rodríguez, M., Nipperess, D. A., Warren, D. L., Laffan, S. W., & VanDerWal, J. (2016). Which species distribution models are more (or less) likely to project broad-scale, climate-induced shifts in species ranges? Ecological Modelling, 342, 135–146. https://doi.org/10.1016/J.ECOLMODEL.2016.10.004

Bernatchez, L. (2016). On the maintenance of genetic variation and adaptation to environmental change: considerations from population genomics in fishes. Journal of Fish Biology, 89(6), 2519–2556. https://doi.org/10.1111/jfb.13145

Branch, C. L., Semenov, G. A., Wagner, D. N., Sonnenberg, B. R., Pitera, A. M., Bridge, E. S., Taylor, S. A., & Pravosudov, V. V. (2022). The genetic basis of spatial cognitive variation in a food-caching bird. Current Biology, 32(1), 210-219.e4. https://doi.org/10.1016/J.CUB.2021.10.036

Breiman, L. (2001). Random Forests. Machine Learning 2001 45:1, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Broennimann, O., Petitpierre, B., Chevalier, M., González-Suárez, M., Jeschke, J. M., Rolland, J., Gray, S. M., Bacher, S., & Guisan, A. (2021). Distance to native climatic niche margins explains establishment success of alien mammals. Nature Communications, 12(1), 1–8. https://doi.org/10.1038/s41467-021-22693-0

Browne, L., Wright, J. W., Fitz-Gibbon, S., Gugger, P. F., & Sork, V. L. (2019). Adaptational lag to temperature in valley oak (Quercus lobata) can be mitigated by genome-informed assisted gene flow. Proceedings of the National Academy of Sciences of the United States of America, 116(50), 25179–25185. https://doi.org/10.1073/PNAS.1908771116

Capblancq, T., Fitzpatrick, M. C., Bay, R. A., Exposito-Alonso, M., & Keller, S. R. (2020). Genomic Prediction of (Mal)Adaptation across Current and Future Climatic Landscapes.

Annual Review of Ecology, Evolution, and Systematics, 51, 245–269. https://doi.org/10.1146/annurev-ecolsys-020720-042553

Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L. P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., Mushayahama, T., … Elser, J. (2021). The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Research, 49(D1), D325–D334. https://doi.org/10.1093/NAR/GKAA1113

Chen, I.-C., Hill, J. K., Ohlemüller, R., Roy, D. B., & Thomas, C. D. (2011). Rapid Range Shifts of Species Associated with High Levels of Climate Warming. Science, 333(6045), 1024–1026. https://doi.org/10.1126/SCIENCE.1206432

Chu, X., Gugger, P. F., Li, L., Zhao, J. L., & Li, Q. J. (2021). Responses of an endemic species (Roscoea humeana) in the Hengduan Mountains to climate change. Diversity and Distributions, 27(11), 2231–2244. https://doi.org/10.1111/DDI.13397

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & Group, 1000 Genomes Project Analysis. (2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158. https://doi.org/10.1093/BIOINFORMATICS/BTR330

Dauphin, B., Rellstab, C., Schmid, M., Zoller, S., Karger, D. N., Brodbeck, S., Guillaume, F., & Gugerli, F. (2021). Genomic vulnerability to rapid climate warming in a tree species with a long generation time. Global Change Biology, 27(6), 1181–1195. https://doi.org/10.1111/GCB.15469

Dawson, T. P., Jackson, S. T., House, J. I., Prentice, I. C., & Mace, G. M. (2011). Beyond Predictions: Biodiversity Conservation in a Changing Climate Downloaded from. Science, 332, 53–58. http://science.sciencemag.org/

Do, C., Waples, R. S., Peel, D., Macbeth, G. M., Tillett, B. J., & Ovenden, J. R. (2014). NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (Ne) from genetic data. Molecular Ecology Resources, 14(1), 209–214. https://doi.org/10.1111/1755-0998.12157

Drovetski, S. V., Zink, R. M., & Mode, N. A. (2009). Patchy distributions belie morphological and genetic homogeneity in rosy-finches. Molecular Phylogenetics and Evolution, 50(3), 437–445. https://doi.org/10.1016/J.YMPEV.2008.12.002

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. Journal of Animal Ecology, 77(4), 802–813. https://doi.org/10.1111/J.1365-2656.2008.01390.X

Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: calculating importance gradients on physical predictors. Ecology, 93(1), 156–168. https://doi.org/10.1890/11-0252.1

Exposito-Alonso, M., Vasseur, F., Ding, W., Wang, G., Burbano, H. A., & Weigel, D. (2017). Genomic basis and evolutionary potential for extreme drought adaptation in Arabidopsis thaliana. Nature Ecology & Evolution, 2(2), 352–358. https://doi.org/10.1038/s41559-017-0423-0

Fedy, B. C., Martin, K., Ritland, C., & Young, J. (2008). Genetic and ecological data provide incongruent interpretations of population structure and dispersal in naturally subdivided populations of white-tailed ptarmigan (Lagopus leucura). Molecular Ecology, 17(8), 1905–1917. https://doi.org/10.1111/J.1365-294X.2008.03720.X

Fitzpatrick, M. C., Chhatre, V. E., Soolanayakanahally, R. Y., & Keller, S. R. (2021). Experimental support for genomic prediction of climate maladaptation using the machine

learning approach Gradient Forests. Molecular Ecology Resources, 00, 1–17. https://doi.org/10.1111/1755-0998.13374
Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. Ecology Letters, 18(1), 1–16. https://doi.org/10.1111/ele.12376
Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Molecular Ecology, 27(9), 2215–2233. https://doi.org/10.1111/MEC.14584
François, O., Martins, H., Caye, K., & Schoville, S. D. (2016). Controlling false discoveries in genome scans for selection. Molecular Ecology, 25(2), 454–469. https://doi.org/10.1111/MEC.13513
Freeman, B. G., Scholer, M. N., Ruiz-Gutierrez, V., & Fitzpatrick, J. W. (2018). Climate change causes upslope shifts and mountaintop extirpations in a tropical bird community. Proceedings of the National Academy of Sciences, 115(47), 11982–11987. https://doi.org/10.1073/PNAS.1804224115
Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution, 6(8), 925–929. https://doi.org/10.1111/2041-210X.12382
Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. Genetics, 196(4), 973–983. https://doi.org/10.1534/GENETICS.113.160572/-/DC1
Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. The Annals of Statistics, 1–67. https://doi.org/10.1214/aos/1176347969
Funk, E. R., Spellman, G. M., Winker, K., Withrow, J. J., Ruegg, K. C., Zavaleta, E., & Taylor, S. A. (2021). Phylogenomic Data Reveal Widespread Introgression Across the Range of an Alpine and Arctic Specialist. Systematic Biology, 70(3), 527–541. https://doi.org/10.1093/sysbio/syaa071
Funk, W. C., Forester, B. R., Converse, S. J., Darst, C., & Morey, S. (2019). Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists. Conservation Genetics, 20(1), 115–134. https://doi.org/10.1007/s10592-018-1096-1
Gougherty, A. V., Keller, S. R., & Fitzpatrick, M. C. (2021). Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. Nature Climate Change, 11(2), 166–171. https://doi.org/10.1038/s41558-020-00968-6
Guisan, A., & Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. Ecology Letters, 8(9), 993–1009. https://doi.org/10.1111/J.1461-0248.2005.00792.X
Hämälä, T., & Savolainen, O. (2019). Genomic Patterns of Local Adaptation under Gene Flow in Arabidopsis lyrata. Molecular Biology and Evolution, 36(11), 2557–2571. https://doi.org/10.1093/MOLBEV/MSZ149
Hastie, T., Tibshirani, R., & Friedman, J. (2009). Random Forests. In The elements of statistical learning: Data mining, inference, and prediction (2nd ed., pp. 587–603). Springer.
Hendricks, D. P. (1977). Brown-Capped Rosy Finch Nesting in New Mexico. The Auk, 94(2), 384–385. https://doi.org/10.1093/AUK/94.2.384
Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., & Whitlock, M. C. (2016). Finding the Genomic Basis of Local

Adaptation: Pitfalls, Practical Solutions, and Future Directions. Https://Doi.Org/10.1086/688018, 188(4), 379–397. https://doi.org/10.1086/688018
Hoffmann, A. A., & Sgró, C. M. (2011). Climate change and evolutionary adaptation. Nature, 470(7335), 479–485. https://doi.org/10.1038/nature09670
Inouye, D. W. (2000). The ecological and evolutionary significance of frost in the context of climate change. Ecology Letters, 3(5), 457–463. https://doi.org/10.1046/J.1461-0248.2000.00165.X
IPCC. (2021). Climate Change 2021: The Physical Science Basis. Contribution of Working Group 1 to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.
Jiménez-Valverde, A., & Lobo, J. M. (2007). Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica, 31(3), 361–369. https://doi.org/10.1016/J.ACTAO.2007.02.001
Johnson, R. E., Hendricks, P., Pattie, D. L., & Hunter, K. B. (2020). Brown-capped Rosy-Finch (Leucosticte australis). In A. F. Poole & F. B. Gill (Eds.), Birds of the World. Cornell Lab of Ornithology. https://doi.org/10.2173/BOW.BCRFIN.01
Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics, 24(11), 1403–1405. https://doi.org/10.1093/BIOINFORMATICS/BTN129
Jonas, T., Rixen, C., Sturm, M., & Stoeckli, V. (2008). How alpine plant growth is linked to snow cover and climate variability. Journal of Geophysical Research: Biogeosciences, 113(G3), 3013. https://doi.org/10.1029/2007JG000680
Kawecki, T., & Ebert, D. (2004). Conceptual issues in local adaptation. Ecology Letters, 7(12), 1225-1241. https://doi.org/10.1111/j.1461-0248.2004.00684.x
Keller, F., Goyette, S., & Beniston, M. (2005). Sensitivity Analysis of Snow Cover to Climate Change Scenarios and Their Impact on Plant Habitats in Alpine Terrain. Climatic Change, 72(3), 299–319. https://doi.org/10.1007/S10584-005-5360-2
Láruson, Á. J., Fitzpatrick, M. C., Keller, S. R., Haller, B. C., & Lotterhos, K. E. (2022). Seeing the forest for the trees: Assessing genetic offset predictions from gradient forest. Evolutionary Applications, 15(3), 403–416. https://doi.org/10.1111/EVA.13354
Lasky, J. R., Forester, B. R., & Reimherr, M. (2018). Coherent synthesis of genomic associations with phenotypes and home environments. Molecular Ecology Resources, 18(1), 91–106. https://doi.org/10.1111/1755-0998.12714
Lek, S., & Guégan, J. F. (1999). Artificial neural networks as a tool in ecological modelling, an introduction. Ecological Modelling, 120(2–3), 65–73. https://doi.org/10.1016/S0304-3800(99)00092-7
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. https://doi.org/10.1093/BIOINFORMATICS/BTR509
Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., & Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. Bioinformatics, 26(22), 2867–2873. https://doi.org/10.1093/BIOINFORMATICS/BTQ559

Martin, A., Zim, H., & Nelson, A. (1961). American wildlife & plants: a guide to wildlife food habits: the use of trees, shrubs, weeds, and herbs by birds and mammals of the United States. Dover Publications Inc.

McCullagh, P., & Nelder, J. (2019). Generalized linear models. https://www.taylorfrancis.com/books/mono/10.1201/9780203753736/generalized-linear-models-mccullagh-nelder

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data.

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Research, 47(D1), D419–D426. https://doi.org/10.1093/NAR/GKY1038

Millar, C. I., Westfall, R. D., & Delany, D. L. (2018). Thermal Components of American Pika Habitat—How does a Small Lagomorph Encounter Climate? Arctic, Antarctic, and Alpine Research, 48(2), 327–343. https://doi.org/10.1657/AAAR0015-046

Neuner, G. (2014). Frost resistance in alpine woody plants. Frontiers in Plant Science, 5(DEC), 654. https://doi.org/10.3389/FPLS.2014.00654/BIBTEX

Nielsen, E. S., Henriques, R., Beger, M., & Heyden, S. von der. (2021). Distinct interspecific and intraspecific vulnerability of coastal species to global change. Global Change Biology, 27(15), 3415–3431. https://doi.org/10.1111/GCB.15651

O'Neill, B. C., Tebaldi, C., Van Vuuren, D. P., Eyring, V., Friedlingstein, P., Hurtt, G., Knutti, R., Kriegler, E., Lamarque, J. F., Lowe, J., Meehl, G. A., Moss, R., Riahi, K., & Sanderson, B. M. (2016). The Scenario Model Intercomparison Project (ScenarioMIP) for CMIP6. Geoscientific Model Development, 9(9), 3461–3482. https://doi.org/10.5194/GMD-9-3461-2016

Oksanen, J., Blanchet, F., Kindt, R., Legendre, P., Minchin, P., O'hara, R. B., Simpson, G. L., Solymos, P., Stevens, M., Wagner, H., & Oksanen, M. (2013). Package "vegan." In Community ecology package (pp. 1–295).

Pacifici, M., Foden, W. B., Visconti, P., Watson, J. E. M., Butchart, S. H. M., Kovacs, K. M., Scheffers, B. R., Hole, D. G., Martin, T. G., Akçakaya, H. R., Corlett, R. T., Huntley, B., Bickford, D., Carr, J. A., Hoffmann, A. A., Midgley, G. F., Pearce-Kelly, P., Pearson, R. G., Williams, S. E., … Rondinini, C. (2015). Assessing species vulnerability to climate change. Nature Climate Change, 5(3), 215–224. https://doi.org/10.1038/nclimate2448

Packard, F. (1968). Brown-capped Rosy Finch. In O. Austin (Ed.), Life histories of North American cardinals, grosbeaks, buntings, towhees, finches, sparrows, and allies (pp. 373–383).

Parmesan, C., & Yohe, G. (2003). A globally coherent fingerprint of climate change impacts across natural systems. Nature, 421(6918), 37–42. https://doi.org/10.1038/nature01286

Patterson, N., Price, A. L., & Reich, D. (2006). Population Structure and Eigenanalysis. PLOS Genetics, 2(12), e190. https://doi.org/10.1371/JOURNAL.PGEN.0020190

Pederson, G. T., Gray, S. T., Ault, T., Marsh, W., Fagre, D. B., Bunn, A. G., Woodhouse, C. A., & Graumlich, L. J. (2011). Climatic Controls on the Snowmelt Hydrology of the Northern

Rocky Mountains. Journal of Climate, 24(6), 1666–1687. https://doi.org/10.1175/2010JCLI3729.1
Pederson, G. T., Gray, S. T., Woodhouse, C. A., Betancourt, J. L., Fagre, D. B., Littell, J. S., Watson, E., Luckman, B. H., & Graumlich, L. J. (2011). The unusual nature of recent snowpack declines in the North American Cordillera. Science, 333(6040), 332–335. https://doi.org/10.1126/SCIENCE.1201570/SUPPL_FILE/PEDERSON-SOM.PDF
Pepin, N. C., Arnone, E., Gobiet, A., Haslinger, K., Kotlarski, S., Notarnicola, C., Palazzi, E., Seibert, P., Serafin, S., Schöner, W., Terzago, S., Thornton, J. M., Vuille, M., Adler, C., Kotlarski, K., & Notarnicola, S. (2022). Climate Changes and Their Elevational Patterns in the Mountains of the World. Reviews of Geophysics, 60(1), e2020RG000730. https://doi.org/10.1029/2020RG000730
Phillips, S. B., Aneja, V. P., Kang, D., & Arya, S. P. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3–4), 231–259. https://doi.org/10.1016/J.ECOLMODEL.2005.03.026
Project, A. (2021). Gridded current and projected climate data for North America at 1 km resolution, generated using the ClimateNA v7.01 software (T. Wang et al., 2021).
Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/BIOINFORMATICS/BTQ033
Rellstab, C., Dauphin, B., & Exposito-Alonso, M. (2021). Prospects and limitations of genomic offset in conservation management. Evolutionary Applications, 14, 1202–1212. https://doi.org/10.1111/eva.13205
Rellstab, C., Zoller, S., Walthert, L., Lesur, I., Pluess, A. R., Graf, R., Bodénès, C., Sperisen, C., Kremer, A., & Gugerli, F. (2016). Signatures of local adaptation in candidate genes of oaks (Quercus spp.) with respect to present and future climatic conditions. Molecular Ecology, 25(23), 5907–5924. https://doi.org/10.1111/MEC.13889
Rodhouse, T. J., Hovland, M., & Jeffress, M. R. (2017). Variation in subsurface thermal characteristics of microrefuges used by range core and peripheral populations of the American pika (Ochotona princeps). Ecology and Evolution, 7(5), 1514–1526. https://doi.org/10.1002/ECE3.2763
Ruegg, K., Bay, R. A., Anderson, E. C., Saracco, J. F., Harrigan, R. J., Whitfield, M., Paxton, E. H., & Smith, T. B. (2018). Ecological genomics predicts climate vulnerability in an endangered southwestern songbird. Ecology Letters, 21(7), 1085–1096. https://doi.org/10.1111/ele.12977
Savolainen, O., Lascoux, M., & Merilä, J. (2013). Ecological genomics of local adaptation. Nature Reviews Genetics, 14(11), 807–820. https://doi.org/10.1038/nrg3522
Scherrer, D., & Körner, C. (2010). Infra-red thermometry of alpine landscapes challenges climatic warming projections. Global Change Biology, 16(9), 2602–2613. https://doi.org/10.1111/J.1365-2486.2009.02122.X
Sclater, W. (1912). A history of the birds of Colorado. Witherby and Co.
Scridel, D., Brambilla, M., de Zwaan, D. R., Froese, N., Wilson, S., Pedrini, P., & Martin, K. (2021). A genus at risk: Predicted current and future distribution of all three Lagopus species reveal sensitivity to climate change and efficacy of protected areas. Diversity and Distributions, 27(9), 1759–1774. https://doi.org/10.1111/DDI.13366

Seastedt, T. R., & Oldfather, M. F. (2021). Climate Change, Ecosystem Processes and Biological Diversity Responses in High Elevation Communities. Climate, 9(5), 87. https://doi.org/10.3390/CLI9050087

Sekercioglu, C. H., Schneider, S. H., Fay, J. P., & Loarie, S. R. (2008). Climate Change, Elevational Range Shifts, and Bird Extinctions. Conservation Biology, 22(1), 140–150. https://doi.org/10.1111/J.1523-1739.2007.00852.X

Shumate, A., & Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. Bioinformatics, 37(12), 1639–1643. https://doi.org/10.1093/BIOINFORMATICS/BTAA1016

Sillero, N. (2011). What does ecological modelling model? A proposed classification of ecological niche models based on their underlying methods. Ecological Modelling, 222(8), 1343–1346. https://doi.org/10.1016/J.ECOLMODEL.2011.01.018

Smith, S. J., & Ellis, N. (2013). gradientForest: Random Forest functions for the census of marine life synthesis project.

Strimas-Mackey, M., Ligocki, S., Auer, T., & Fink, D. (2021). ebirdst: Tools for loading, plotting, mapping and analysis of eBird STatus and Trends data products (1.0.0).

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. Biological Conservation, 142(10), 2282–2292. https://doi.org/10.1016/J.BIOCON.2009.05.006

Team, R. C. (2019). R: A Language and Environment for Statistical Computing (3.5). R Foundation for Statistical Computing.

Thuiller, W., Georges, D., Engler, R., & Breiner, F. (2016). Package biomod2.

Tigano, A., & Friesen, V. L. (2016). Genomics of local adaptation with gene flow. Molecular Ecology, 25(10), 2144–2164. https://doi.org/10.1111/MEC.13606

Urban, M. C. (2015). Accelerating extinction risk from climate change. Science, 348(6234), 571–573. https://doi.org/10.1126/SCIENCE.AAA4984

de Villemereuil, P., Gaggiotti, O. E., Mouterde, M., & Till-Bottraud, I. (2016). Common garden experiments in the genomic era: new perspectives and opportunities. Heredity, 116, 249-254. https://doi.org/10.1038/hdy.2015.93

Wang, T., Hamann, A., Spittlehouse, D., & Carroll, C. (2016). Locally Downscaled and Spatially Customizable Climate Data for Historical and Future Periods for North America. PLOS ONE, 11(6), e0156720. https://doi.org/10.1371/JOURNAL.PONE.0156720

Warren, E. R. (1916). Notes on the Birds of the Elk Mountain Region, Gunnison County, Colorado. The Auk, 33(3), 292–317. https://doi.org/10.2307/4072327

Yeaman, S. (2015). Local adaptation by alleles of small effect. The American Naturalist, 186, S74–S89. https://doi.org/10.1086/682405/SUPPL_FILE/55813APA.PDF

Zheng, X., Levine, D., Shen, J., Gogarten, S. M., Laurie, C., & Weir, B. S. (2012). A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics, 28(24), 3326–3328. https://doi.org/10.1093/BIOINFORMATICS/BTS606

Zimmerman, S. J., Aldridge, C. L., Langin, K. M., Wann, G. T., Scott Cornman, R., & Oyler-McCance, S. J. (2021). Environmental gradients of selection for an alpine-obligate bird, the white-tailed ptarmigan (Lagopus leucura). Heredity, 126(1), 117–131. https://doi.org/10.1038/s41437-020-0352-6

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1(1), 3–14. https://doi.org/10.1111/J.2041-210X.2009.00001.X

# 2. POPULATION ASSIGNMENT FROM GENOTYPE LIKELIHOODS FOR LOW-COVERAGE SEQUENCING DATA


## Summary

Low-coverage whole genome sequencing (WGS) is increasingly used for the study of evolution and ecology in both model and non-model organisms; however, effective application of low-coverage WGS data requires the implementation of probabilistic frameworks to account for the uncertainties in genotype likelihood data. Here, we present a probabilistic framework for using genotype likelihood data for standard population assignment applications. Additionally, we derive the Fisher information for allele frequency from genotype likelihood data and use that to describe a novel metric, the effective sample size, which figures heavily in assignment accuracy. We make these developments available for application through WGSassign, an open-source software package that is computationally efficient for working with whole genome data. Using simulated and empirical data sets, we demonstrate the behavior of our assignment method across a range of population structures, sample sizes, and read depths. Through these results, we show that WGSassign can provide highly accurate assignment, even for samples with low average read depths (< 0.01X) and among weakly differentiated populations. Our simulation results highlight the importance of equalizing the effective sample sizes among source populations in order to achieve accurate population assignment with low-coverage WGS data. We further provide study design recommendations for population-assignment studies and discuss the broad utility of effective sample size for studies using low-coverage WGS data.

## Introduction

In just a few years, next-generation sequencing (NGS) technologies have revolutionized the study of evolution and ecology in both model and non-model organisms, and have become established as standard tools in molecular ecology. In particular, whole genome sequencing (WGS) can provide sequence data from a large proportion of the genome and is increasing in use. While large-scale WGS projects can be prohibitively expensive at the necessary read depths for accurately calling individual genotypes, low-coverage WGS offers a cost-effective approach aimed at reducing the read depth per individual while retaining sufficient information for genomic analyses. However, since low-coverage WGS precludes the ability to call individual genotypes, probabilistic frameworks are used to account for the uncertainty in an individual's genotype (Nielsen *et al.* 2011; Buerkle & Gompert 2013). Extending common analyses in the field of molecular ecology to accommodate genotype uncertainty through the direct use of genotype likelihoods is a necessary advance for broadening the utility of low-coverage WGS.

The creation of probabilistic frameworks for allele frequency estimation, genotype calling, and single nucleotide polymorphism (SNP) calling have made low-coverage WGS practical for many applications (Nielsen *et al.* 2011, 2012; Kim *et al.* 2011). By first estimating the joint site frequency spectrum for individuals without calling individual genotypes, priors on allele frequency can improve the calling of individuals' genotypes and SNPs. Population genetics analyses have been further advanced through the development of methods that quantify genetic differentiation and investigate population structure with principal components analysis, while accounting for uncertain genotypes (Fumagalli *et al.* 2013). Similarly, accurate estimates of individual admixture proportions (Skotte *et al.* 2013) and pairwise relatedness (Korneliussen & Moltke 2015) can be obtained using genotype likelihoods. The widespread use of these

methods is facilitated by software that is both user-friendly and computationally efficient (e.g. ANGSD (Korneliussen *et al.* 2014), ngsTools (Fumagalli *et al.* 2014), PCangsd (Meisner & Albrechtsen 2018)). However, a fundamental analysis for molecular ecology yet to be developed for low-coverage WGS data is population assignment.

Population assignment methods are used to determine an individual's population of origin and have provided insight into ecological and evolutionary processes, such as dispersal, hybridization, and migration, as well as informed conservation and management decisions (Manel *et al.* 2005). The traditional assignment test uses an individual's multilocus genotype and the source populations' allele frequencies to calculate the likelihood of the genotype originating from each of the populations (Paetkau *et al.* 1995; Rannala & Mountain 1997). Using this framework, the recent increase in available markers (e.g., from RADseq approaches) has made possible highly accurate assignment of individuals among weakly differentiated populations by using subsets of informative loci for population structure (e.g. DeSaix *et al.* 2019; Ruegg *et al.* 2014; Benestan *et al.* 2015). The traditional assignment test is readily extended to analyses such as genetic stock identification (GSI), to determine the proportion of source populations in a mixture of individuals Smouse *et al.* (1990). To date, methods for performing assignment tests require known genotypes and have not been implemented to use genotype likelihoods.

Assignment tests are well suited for application with low-coverage WGS data, because they rely heavily on allele frequency estimates, for which a number of approaches are already developed. For accurate allele frequency estimation from low coverage WGS data, simulation studies have demonstrated that prioritizing larger sample sizes of individuals with lower sequencing depth is the most cost-effective strategy (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013). Specific recommendations include aiming for individual sequencing depths of

1x (Buerkle & Gompert 2013) or having at least 10 individuals sequenced with a total per-population sequencing depth of at least 10x (Lou *et al.* 2021). The goal of these strategies is to maximize information for estimating allele frequencies given finite resources for sequencing depth and number of samples. Lower sequencing depth decreases the amount of information about population allele frequency, while using larger sample sizes increases the amount of information. However, information is not directly quantified in these studies; rather comparison of known versus simulated allele frequencies were used to arrive at these general rules of thumb (Buerkle & Gompert 2013; Lou *et al.* 2021). The development of an information metric that accounts for read-depth variation across genotypes would provide a valuable method to quantify the thresholds of information needed for parameter estimation with low-coverage WGS data.

Here we present WGSassign, an open-source software package of population assignment tools for genotype likelihood data from low coverage WGS. The objectives of WGSassign are: 1) provide common assignment methods that use genotype likelihoods, instead of called genotypes, 2) evaluate the information available in low-read-depth sequencing data for allele frequency estimation, and 3) achieve computational efficiency for processing large numbers of samples with genome-wide data. WGSassign provides methods for individual assignment, estimation of mixture proportions, and leave-one-out cross-validation of samples of known origin. Additionally, it calculates a *z*-score metric that can indicate when samples originate from an unsampled source population. For the second objective, we calculate Fisher Information and determine the *effective sample size*—the number of samples with completely observed genotypes that would yield the same amount of statistical information for estimating allele frequency as the observed genotype likelihoods in a dataset. This calculation of effective sample size has broad utility for population genomics studies using low-coverage WGS.

We validate WGSassign and investigate its behavior with an extensive set of simulations and demonstrate its use on two empirical datasets. In the first, we apply WGSassign to weakly differentiated groups of yellow warblers (*Setophagia petechia*). In the second, we apply WGSassign to two well-differentiated Chinook salmon (*Oncorhynchus tshawytscha*) populations to demonstrate that when sufficient effective sample sizes of the source population are available, unknown individuals can be assigned accurately, even at extremely low read depths.

## Methods

WGSassign is written in Python 3 (https://www.python.org/) and requires the following modules: numpy (https://numpy.org/), cython (https://cython.org/), and scipy (https://scipy.org/). Detailed instructions for using WGSassign are available at https://github.com/mgdesaix/WGSassign.

### *Population Assignment*

We assume that there are $K$ sampled source populations to which an individual can be assigned using data from $L$ biallelic loci in the genome. Let a diploid individual's genotype at locus $\ell$ ($1 \leq \ell \leq L$) be represented by $G_\ell \in \{0, 1, 2\}$, which counts the number of alleles matching the reference genome carried by the individual at locus $\ell$. Denote by $\theta_{k,\ell}$ the true—but typically unkown—frequency of the alternate allele at locus $\ell$ within source population $k$. Under the assumption of Hardy-Weinberg equilibrium, the probability of $G_\ell$, when the individual is from population $k$ is:

$$P(G_\ell|\theta_{k,\ell}) = \begin{cases} (1-\theta_{k,\ell})^2 & \text{if } G_\ell = 0 \\ 2(\theta_{k,\ell})(1-\theta_{k,l}) & \text{if } G_\ell = 1 \\ (\theta_{k,\ell})^2 & \text{if } G_\ell = 2. \end{cases} \quad (1)$$

With low-coverage sequencing data, Gℓ is not observed with certainty. Rather, evidence about the unknown genotype is obtained from sequencing reads covering the locus. Let Rℓ denote the sequencing read data from an individual at locus ℓ. The evidence for the state of Gℓ from the read data is summarized as the likelihood of the genotype given the read data, which is simply the probability of the read data given the genotype, considered as a function of the genotype:

$$P(R_\ell|G_\ell) = \begin{cases} g_{\ell,0} & \text{for } G_\ell = 0 \\ g_{\ell,1} & \text{for } G_\ell = 1 \\ g_{\ell,2} & \text{for } G_\ell = 2. \end{cases} \quad (2)$$

Without loss of generality, we consider these likelihoods to be scaled so that they sum to one: $g_{\ell,0} + g_{\ell,1} + g_{\ell,2} = 1$. Such likelihoods are typically a function of the number of reads of each allele observed and the corresponding base quality scores, and they are computed during genotype calling by a variety of programs such as bcftools (Li *et al.* 2009; Li 2011), GATK (McKenna *et al.* 2010), and ANGSD (Korneliussen *et al.* 2014). An accessible review of the different models providing genotype likelihoods is found in (Lou *et al.* 2021).

To do population assignment from the read data of an individual (rather than from directly observed genotypes) requires, for each locus, $\ell$, the likelihood that the individual came from a source population $k$, say, given the individual's read data. This is simply the probability of the read data from the individual given that the individual came from source population $k$,

with allele frequencies $\theta_{k,\ell}$. Thus, we require $P(R_\ell|\theta_{k,\ell})$, which can be calculated from (1) and (2) using the law of total probability:

$$\begin{aligned} P(R_\ell|\theta_{k,\ell}) &= \sum_{G_\ell=0}^{2} P(R_\ell|G_\ell)P(G_\ell|\theta_{k,\ell}) \\ &= g_{\ell,0}(1-\theta_{k,\ell})^2 + g_{\ell,1}2(\theta_{k,\ell})(1-\theta_{k,\ell}) + \\ &\quad g_{\ell,2}(\theta_{k,\ell})^2. \end{aligned} \tag{3}$$

If the $L$ loci in the genome are not in linkage disequilibrium (LD), and are hence independent of one another, within source populations, then the likelihood of source population $k$ given $R$, the read sequencing data across the entire genome, is simply the product over loci.

$$P(R|\theta_k) = \prod_{\ell=1}^{L} P(R_\ell|\theta_{k,\ell}), \tag{4}$$

where $\theta_k$ denotes the set of all $L$ allele frequencies in population $k$. Of course, with lcWGS some variants may be near one another and will then likely be in LD. In such a case (4) is not correct, but, rather, is a composite-likelihood approximation to the true likelihood (which is largely intractable). Composite likelihood estimators often produce unbiased results, but, because they do not take account of the dependence of different variables in the likelihood, they typically underestimate the uncertainty in the estimates (Larribe & Fearnhead 2011). For each individual of unknown origin, this likelihood can be computed for each source population, $k$, and the relative values of those likelihoods gives the evidence that the individual came from each of the source populations. If the prior probability $\pi_k$ that an individual came from source population $k$ is available for $k \in \{1, \ldots, K\}$, then the likelihoods can be used to compute the posterior probability that the individual came from each of the source populations:

$$P(Z = k|R, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) = \frac{\pi_k P(R|\theta_k)}{\sum_{i=1}^{K} \pi_k P(R|\theta_k)}, \tag{5}$$

where $Z$ is a random variable indicating the origin of the individual.

In practice, the allele frequencies in each source population are not known with certainty. Accordingly, these frequencies must be estimated from sequencing read data from individuals known to be from the source populations (these are often referred to as "reference samples.") We estimate these by maximum likelihood. The probability of the read data, $R_l^{(i)}$, from the $i^{th}$ reference sample, given that it came from source population $k$, is, following (3),

$$P(R_\ell^{(i)}|\theta_{k,\ell}) = \\ g_{\ell,0}^{(i)}(1-\theta_{k,\ell})^2 + g_{\ell,1}^{(i)}2(\theta_{k,\ell})(1-\theta_{k,\ell}) + g_{\ell,2}^{(i)}(\theta_{k,\ell})^2, \tag{6}$$

where the genotype likelihoods are now adorned with a superscript $^{(i)}$ to denote they are for the $i^{th}$ reference sample. Assuming the samples from source population $k$ are not related, the log-likelihood for $\theta_{k,\ell}$ given the read data from all $n_k$ reference samples from population $k$ is:

$$L(\theta_{k,\ell}) = \sum_{i=1}^{n_k} \log P(R_\ell^{(i)}|\theta_{k,\ell}) \tag{7}$$

In our implementation, we first use the Expectation-Maximization algorithm (Dempster et al. 1977) described in the supplement to Meisner & Albrechtsen (2018) to obtain the maximum likelihood estimates (MLEs) of the population allele frequencies, $\hat{\theta}_{k,l}$, from the reference samples. Then, when calculating $P(R|\theta_k)$ we substitute $\tilde{\theta}_{k,l}$ for $\theta_{k,l}$ calculated as follows:

$$\tilde{\theta}_{k,\ell} = \begin{cases} \hat{\theta}_{k,\ell} & \text{if } \hat{\theta}_{k,\ell} > 0 \\ \frac{1}{2(n_k+1)} & \text{if } \hat{\theta}_{k,\ell} = 0, \\ 1 - \frac{1}{2(n_k+1)} & \text{if } \hat{\theta}_{k,\ell} = 1, \end{cases} \quad (8)$$

where, again, $n_k$ is the number of reference samples from source population *k*. This provides a correction for cases in which the allele exists in a source population, but was not detected in the reference samples from that population—effectively, it adds one more individual to the sample that carries one copy of the allele not previously seen in that reference population. As should be clear from the preceding development, the accuracy of population assignment depends, at least in part, on the accuracy of the estimates of the allele frequencies from each source population. The following section develops theory (which is then implemented in WGSASSIGN) that provides the user with a measure of allele frequency estimate accuracy, calculated from the genotype likelihoods in the reference samples, that takes account of both sample size and read depth.

*Fisher Information and Effective Sample Size*

The likelihood that an individual originated from a source population depends on the read data (summarized as a genotype likelihood) and also on the estimated allele frequencies of the source populations. In turn, the accuracy of the estimated allele frequency depends on the number of individuals in the reference sample from the source population and read depth of those individuals (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013). Fewer individuals sampled and lower sequencing depth will result in less information in the data regarding allele frequency. As noted above, estimates of the allele frequencies are made by maximum likelihood using the sequencing data on the reference samples from each source population. Fisher information is a statistical metric that quantifies the amount of information in

a sample for estimating an unknown, continuous parameter (Fisher 1922). It measures the curvature of the log-likelihood function, and is inversely related to the variance. In visual terms, a sharply peaked log-likelihood curve (i.e., one with greater curvature) for a parameter indicates greater certainty in the estimated parameter (and, also higher Fisher information) than a flatter log-likelihood function. Formally, the curvature is measured by the negative second derivative of the log-likelihood function. The *observed* Fisher information for allele frequency is that negative second derivative evaluated at the MLE

$$I_o(\theta_{k,\ell}) = -\frac{\partial^2 L(\theta_{k,\ell})}{\partial \theta_{k,\ell}^2}\Big|_{\theta_{k,\ell}=\hat{\theta}_{k,\ell}}. \quad (9)$$

The $I_o(\theta_{k,\ell})$, the observed Fisher information for $\theta_{k,\ell}$ in the reads from a single individual, $i$, is found to be:

$$I_o^{(i)}(\theta_{k,\ell}) = \left[\frac{2(g_{\ell,0}^{(i)} + g_{\ell,2}^{(i)} - 2g_{\ell,1}^{(i)})}{g_{\ell,0}^{(i)}(1-\hat{\theta}_{k,\ell})^2 + g_{\ell,1}^{(i)}2\hat{\theta}_{k,\ell}(1-\theta_{k,\ell}) + g_{\ell,2}^{(i)}\hat{\theta}_{k,\ell}^2} + \left(\frac{2\hat{\theta}_{k,\ell}(g_{\ell,0}^{(i)} + g_{\ell,2}^{(i)} - 2g_{\ell,1}^{(i)}) + 2(g_{\ell,1}^{(i)} - g_{\ell,0}^{(i)})}{g_{\ell,0}^{(i)}(1-\hat{\theta}_{k,\ell})^2 + g_{\ell,1}^{(i)}2\hat{\theta}_{k,\ell}(1-\hat{\theta}_{k,\ell}) + g_{i,2}^{(i)}\hat{\theta}_{k,\ell}^2}\right)^2\right]. \quad (10)$$

The observed Fisher information from all $n_k$ reference samples is then simply, $I_o(\theta_{k,\ell}) = \sum i{=}1\ I_o(\theta_{k,\ell})$. To derive $\tilde{n}_l$ our effective sample size metric for locus $\ell$, we compare this observed Fisher information to the *expected* Fisher information that would be obtained from $2\tilde{n}_l$ gene copies with allelic type directly observed from a population in which the true allele frequency is $\hat{\theta}_{k,l}$:

$$I_e(\theta_{k,\ell}) = \frac{2\tilde{n}_\ell}{\hat{\theta}_{k,\ell}(1-\hat{\theta}_{k,\ell})}. \quad (11)$$

Equating $I_o(\theta_{k,\ell})$ to $I_e(\theta_{k,\ell})$ and solving for $\tilde{n}_l$ yields

$$\tilde{n}_\ell = \frac{1}{2} I_o(\theta_{k,\ell}) \times \hat{\theta}_{k,\ell}(1 - \hat{\theta}_{k,\ell}). \tag{12}$$

This is the number of diploid individuals with perfectly observed genotypes that provides the same information (and hence accuracy) for estimating $\theta_{k,\ell}$ as is available from the sequencing read data from the $n_k$ reference samples from source population $k$. We term $\tilde{n}_l$, calculated as above, the *effective sample size* of the read data from the reference samples of source population $k$ at locus $\ell$. In practice, to avoid issues of non-differentiability on the boundaries of the space (i.e., at $\theta = 0$ or $\theta = 1$) we calculate $\tilde{n}_l$ using $\tilde{\theta}_{k,l}$. The effective sample size for an individual is then derived by taking the mean of $\tilde{n}_l$ across all loci, $\tilde{n} = \frac{1}{L}\sum_{l=1}^{L} \tilde{n}_l$

Fisher information and effective sample size calculated in this way are useful summaries for understanding the trade-offs between sequencing more individuals at lower depth versus fewer individuals at higher depth, at least as it pertains to accurately estimating allele frequencies. In the context of population assignment, the effective sample size, in particular, provides an accessible metric for how good (or bad) the source-population allele frequencies can be expected to be. As we will see later, Fisher information also provides a valuable way to standardize the effective sample size of the reference samples from each population—an important consideration when using WGSassign. A useful statistic for accomplishing this is the individual-specific average effective size for individual $i$:

$$\tilde{n}^{(i)} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{1}{2} I_o^{(i)}(\theta_{k,\ell}) \times \hat{\theta}_{k,\ell}(1 - \hat{\theta}_{k,\ell}), \tag{13}$$

Where $I_o^i(\theta_{k,l})$ is the contribution to the observed Fisher information of the reads from individual *i*

$$I_o^{(i)}(\theta_{k,\ell}) = -\frac{\partial^2 \log P(R_\ell^{(i)}|\theta_{k,\ell})}{\partial \theta_{k,\ell}^2}\Big|_{\theta_{k,\ell}=\hat{\theta}_{k,\ell}}. \quad (14)$$

$\tilde{n}^{(i)}$ ranges between 0 and 1.

We also implement a *z*-score calculation for determining whether an individual's genotype is unlikely to have come from one of the *K* source populations, but rather, from an unsampled population. In short, we determine the expected distribution of log probabilities of an individual's genotype likelihood data arising from a population (given the individual's allele counts across loci and the population's allele frequencies), using a central limit theorem approximation. The *z*-score is then calculated by subtracting the mean expected likelihood from the observed likelihood and dividing the difference by the standard deviation of the expected likelihoods. Given that the actual distribution of the *z*-score is likely to deviate from a standard normal distribution, we further standardize the observed *z*-score by the *z*-scores of the reference individuals from the source populations. Individuals truly from an assigned population are expected to have *z*-scores within several standard deviations of the normal distribution, while individuals from an unsampled but differentiated population are expected to have *z*-scores that fall below the expected range of a standard unit normal random variate.

*Simulations to illustrate the effective sample size*

We used the R programming language to run simulations that illustrate how Fisher information and effective sample size vary across a range of simulated read depths and true allele frequencies. Our simulations assumed a sample size of 100 diploid individuals and a single

biallelic locus, with allelic types within individuals being independent of each other. For each individual, we simulated read depth from a Poisson distribution with mean $D_{ave}$ and allelic types upon each read by sampling from the two gene copies within the individual with equal probability and switching the allelic type with probability 0.01 for each read to simulate sequencing errors. Genotype likelihoods from the reads were calculated according to the simulation model. We calculated the maximum likelihood estimate (MLE) for $\theta$ from the genotype data as the observed proportion of alleles, and for the sequencing read data, we used the EM algorithm to compute the MLE. Using these estimates, we then computed the observed information from the genotypes and from the genotype likelihoods. To determine the effective sample size, we calculated the expected information for observed genotypes, assuming the true value of $\theta$ was the MLE from genotype likelihoods and then used (12). We ran these simulations across values of $D_{ave} \in \{0.1, 0.5, 1, 2, 3, 4, 5, 7, 10, 15, 20, 30, 50\}$ and values of $\theta \in \{0.01, 0.05, 0.10, \ldots, 0.90, 0.95, 0.99\}$, simulating 50 replicate samples for each combination.

*Genetic Simulations*

To demonstrate the efficacy of WGSassign in performing population assignment for a range of samples, read depths and genetic differentiation among populations we simulated a series of genetic datasets using msprime (Kelleher *et al.* 2016). In the first simulation, we implemented two-population island models with an effective population size of 1000 individuals in each population. We simulated ancestry for a genomic sequence of 108 bases with a recombination rate of 10−8 and a mutation rate of 10−7. To vary the genetic differentiation between populations, we varied the lineage migration rate parameter between 0.0005 and 0.05 in 20 equal increments. From both populations we sampled 10, 50, 100, or 500 individuals.

Pairwise $F_{ST}$ was calculated between the two populations using the sampled individuals and the genetic variants were output in variant call format. Genotype likelihoods were produced with vcfgl (https://github.com/isinaltinkaya/vcfgl) based on mean read depths of 0.1X, 0.5X, 1X, 5X, 10X, or 50X. For each of the 480 parameter combinations (10 migration rates, 4 sample sizes, and 6 read depths) we simulated 10 replicates, for a total of 2,400 simulated datasets. We used bcftools (Li *et al.* 2009; Li 2011) to remove any SNPs with a minor allele frequency less than 0.05. We converted the data to Beagle file format with custom scripts, and used these data as input into WGSassign.

To determine the influence of sampling design (i.e. number of samples in a source population and their read depths), as well as amount of genetic differentiation, on assignment accuracy, we calculated the effective sample size and leave-one-out (LOO) assignment accuracy for each population. In WGSassign, LOO is performed by iteratively removing an individual of known origin from its source population, calculating allele frequencies within the source populations using the remaining individuals, and then calculating the likelihood that the removed individuals originated from each of the different source populations. The LOO method is widely used to avoid the bias that arises from using training data that also includes data being tested. The assigned population was determined by maximum likelihood. We also measured the run time for the calculation of allele frequency and effective sample size, as well as the LOO calculation.

In the second simulation, we assessed the influence on assignment accuracy of using unequal effective sample sizes of source populations. In population assignment applications, unequal sample sizes in different populations will result in different levels of precision in the allele frequency estimation. We implemented two-population island models as in the previous

simulation, but included all sample combinations of 10, 50, and 100 individuals for the two populations. We also used 10 equal increments of migration rates from 0.005 to 0.05, and simulated read depths of 1X, 5X, and 10X. We then filtered by a minor allele frequency of 0.05 and randomly selected 100,000 SNPs to be used for the effective sample size calculation and LOO assignment.

In the third simulation, we assessed the performance of the WGSassign *z*-score metric for determining whether an individual of unknown origin being assigned to a population is actually from an unsampled population. We implemented a three-population stepping-stone model with 20, 60, or 110 individuals using msprime. Individuals had simulated mean read depths of 1X or 5X, and we customized vcfgl (https://github.com/isinaltinkaya/vcfgl) to output allele counts for the major and minor alleles. We used populations 1 and 2 in the stepping-stone model as reference populations and calculated the reference *z*-scores using WGSassign from all but 10 of the individuals in these two populations. We assigned 10 individuals from population 3 and 10 from population 2 to the reference populations (1 and 2) using WGSassign. We calculated the *z*-scores of these individuals' assignments to demonstrate the behavior of the *z*-score metric for correctly assigned individuals (i.e., the individuals from population 2 that were assigned to population 2) versus individuals from an unsampled population (i.e., the individuals from population 3 that were assigned to population 2).

*Application to Empirical Data*

We used WGSassign on data from yellow warblers to test its accuracy when applied to individuals from a species exhibiting isolation by distance (Bay *et al.* 2021; Gibbs *et al.* 2000). Previous work on yellow warblers has found weak differentiation between populations, with

pairwise $F_{ST}$ values on the order of 0.01 or less (Gibbs *et al.* 2000). Blood samples from 105 individuals was collected via brachial venipuncture in the years 2020 and 2021. These served as reference samples from 3 populations—North, Central, and South—previously described in Bay *et al.* (2021) and Gibbs *et al.* (2000). We extracted DNA from blood using the manufacturer's protocol for Qiagen DNEasy Blood and Tissue Kits. Whole genome sequencing libraries were prepared following modifications of Illumina's Nextera Library Preparation protocol (Schweizer & DeSaix 2023) and sequenced on a HiSeq 4000 at Novogene Corporation Inc., with a target sequencing depth of 2X per individual.

Sequences were trimmed with TrimGalore version 0.6.5 (https://github.com/FelixKrueger/TrimGalore) and mapped to the NCBI yellow warbler reference genome (Sayers *et al.* 2022) (accession number JANCRA010000000) using the Burrows-Wheeler Aligner software version 0.7.17 (Li & Durbin 2009). After mapping, the resulting SAM files were sorted, converted to BAM files, and indexed using Samtools version 1.9 (Li *et al.* 2009). We used MarkDuplicates from GATK version 4.1.4.0 (McKenna *et al.* 2010) to mark read duplicates and clipped overlapping reads with the clipOverlap function from bamUtil (https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap). To reduce sequencing depth variation, we used the DownsampleSam function from GATK to down sample reads from BAM files with greater than 2X coverage, to 2X coverage. To identify genetic markers from low-coverage WGS data, we used stringent filtering options in ANGSD version 0.9.40 (Korneliussen *et al.* 2014). We retained reads with a mapping quality of at least 30 and base quality of at least 33. We retained SNPs that had read data in at least 50% of individuals and a minor allele frequency greater than 0.05. The filtered variants were output as genotype likelihoods and stored in a Beagle-formatted file.

We implemented principal components analysis (PCA) to ensure reference samples from each of our source populations actually showed geographic signatures of clustering in the PCA. Genetic differentiation among the breeding populations was calculated by creating site allele frequency files for each breeding population and calculating $F_{ST}$ in ANGSD (Korneliussen *et al.* 2014). In order to assess our ability to accurately assign individuals of unknown origin to breeding populations, we determined the accuracy of assignment of the known breeding origin individuals using WGSassign's leave-one-out approach.

For the second empirical dataset, we applied WGSassign to previously published data from Chinook salmon (Thompson *et al.* 2020) to assess its utility in situations with low to extremely low read depth and poor-quality DNA. For this scenario, we entertained the task of assigning Chinook salmon to either the Klamath River basin, or the Sacramento Basin. These populations are quite distinct, with pairwise $F_{ST}$ values between the basins on the order of 0.1. So, it should be quite easy to distinguish fish from the two basins. However, in whole genome sequencing data from Thompson *et al.* (2020) there were several fish from rivers in the Klamath basin collected from carcasses with low read depth. These fish were excluded from most analyses in Thompson *et al.* (2020) because they did not reliably cluster with other fish from their populations on a PCA; however we evaluate here if their basin of origin can be recovered using WGSassign. Additionally, through downsampling of reads from the BAM files we investigate if average read depths as low as 0.001X in the sample being assigned can deliver accurate assignments.

We included fish from the closely related Feather River Spring, Feather River Fall, San Joaquin Fall, and Coleman Late Fall collections as members of the Sacramento River source population, while fish from the closely related Salmon River Fall and Spring and Trinity River

Fall and Spring collections constitute samples from the Klamath River source population. With 64 fish in each source population, we removed the 12 fish from each that had the fewest sequencing reads to serve as our 24 "unknown" fish to be assigned to the populations. The remaining 52 in each population served as the reference samples.

The genotype likelihoods for the reference sample were in a VCF file produced by GATK. This was filtered using bcftools (Danecek *et al.* 2021) to retain only biallelic SNPs with a minor allele frequency > 0.05 which were missing data in fewer than 30% of the samples. Additionally, data from chromosome 28, which holds a region strongly differentiated between spring-run and fall-run Chinook salmon (Thompson *et al.* 2020) was excluded. These genotype likelihoods were stored in a Beagle-formatted file using a custom script.

The data for the test samples were extracted from BAM files. We used samtools stats (Li *et al.* 2009) to determine the average read depth in each BAM and used that number with samtools view to downsample each BAM five times with five separate seeds to average read depth levels of 0.001X, 0.005X, 0.01X, 0.05X, 0.1X, 0.5X, and 1.0X, when those read depths were lower than the full read depth of the file. Genotype likelihoods for the 24 individuals were then called with ANGSD v0.940 (Korneliussen *et al.* 2014) using the -sites options to call only the sites found in the Beagle-formatted file of the reference samples. After genotype likelihood estimation in the test samples, the Beagle file of reference samples was filtered to include only the sites output by ANGSD. The resulting Beagle files were then passed to WGSassign to compute the likelihood of population origin for each of the test fish, and the results were plotted using R version 4.0 (R Core Team 2022).

## Results

*Effective Sample Size Simulations*

As expected, observed Fisher information for allele frequency from sequencing read data increases as the average sequencing depth increases, reaching a limit at the observed information from fully observed genotypes. The absolute value of the observed Fisher information varies widely over the different allele frequencies, however the relative values of information from genotypes and from sequencing reads varies less, and the effective sample size is largely consistent across the range of minor allele frequencies from 0.05 to 0.5, showing the effective sample size to be a useful metric. Fisher information and effective sample size are shown for three representative values of $\theta$ (0.05, 0.3, and 0.5) in Figure 2.1. The flattening of the curves for observed information from sequencing data as the average read depth increases indicates the diminishing returns of additional sequencing depth versus additional samples, for estimating allele frequencies that has been noted previously (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013).

*Genetic Simulations*

In the first simulation, genetic differentiation between the sampled individuals from the two populations ranged from -0.003 - 0.13 $F_{ST}$. Across all read depths within each category of number of samples (10, 50, 100, 500), assignment accuracy increased with genetic differentiation, and generally high assignment accuracy was achieved even with low genetic differentiation (Figure 2.2). Accuracy above 90% was reached for all simulations within the 500 samples category with $F_{ST} > 0.004$, 100 samples category with $F_{ST} > 0.006$, 50 samples category with $F_{ST} > 0.015$, and the 10 samples category with $F_{ST} > 0.043$. When excluding simulations

with populations with the lowest effective sample sizes (< 0.1 individuals), high assignment accuracy was reached for all simulations at $F_{ST} > 0.015$ (Figure 2.2). Within each sample size category, increasing average read depth, and therefore effective sample size, resulted in higher assignment accuracy, especially when populations had weak genetic differentiation (Figure 2.2).

Runtime for the simultaneous calculation of Fisher information, effective sample size, and allele frequency for populations in WGSassign was fast. With 2 populations and 100,000 loci being analyzed in parallel with 20 threads, runtime was less than 10 seconds for populations with 100 samples or less, and between 15 and 30 seconds for populations with 500 samples. Leave-one-out assignment requires population allele frequency to be recalculated for each individual in the population, and time required for that re-calculation increases linearly with sample size. Accordingly, runtime for LOO cross-validation is expected to increase quadratically with increasing number of samples per population, and we observe this: for 100 samples for the two populations at 1X mean individual read depth LOO assignment had a mean runtime of 51 seconds and for 500 samples run time was 1,743 seconds. Run times also increase with lower read depth due to the increase in iterations needed in the expectation-maximization algorithm for allele frequency calculation used from PCangsd (Meisner & Albrechtsen 2018).

When $F_{ST}$ is greater than 0.01, effective sample sizes as low as approximately 3 individuals achieve assignment accuracy of greater than 90% (Figure 2.3). Examining simulations with weak genetic differentiation ($0.005 < F_{ST} < 0.01$), shows that a minimum effective sample size of 10 individuals is needed for consistently high assignment accuracy (Figure 2.3). At the weakest genetic differentiation of $F_{ST} < 0.005$, consistently high assignment accuracy is not necessarily achieved across all simulations, but a minimum effective sample size of 100 individuals is needed for an assignment accuracy of greater than 80%.

*Assignment bias due to unequal sample sizes*

Our simulation results for unequal sample sizes demonstrate that high assignment bias occurs when populations have different numbers of samples (Figure 2.4). When populations have the same number of samples, with the same average read depths, assignment accuracy overall increases with genetic differentiation and there is no evidence of bias, with one population having higher accuracy than another population. However, when populations have unequal sample sizes, individuals from the less-sampled population tend to be assigned to the more-sampled population, even when genetic differentiation is higher ($F_{ST} > 0.01$). This bias is exacerbated when effective sample size is lower (i.e. the populations have lower read depths).

*Determining an individual's origin from an unsampled population*

At higher genetic differentiation ($F_{ST} > 0.1$), samples can readily be identified as coming from an unsampled population using the *z*-score metric in WGSassign (Figure 2.5). At such high differentiation, individuals from an unsampled population tend to have *z*-scores less than 3 compared to individuals correctly assigned to a population having *z*-scores in (−3, 3), as expected of a standard unit normal. With weaker genetic differentiation ($F_{ST} < 0.1$), sample size and read depth have a more noticeable effect on the behavior of the *z*-score metric (Figure 2.5). Generally, higher source sample sizes and read depths allow individuals from unsampled populations to be distinctively identified from individuals that are truly from a source population.

*Application to Empirical Data*

Yellow warbler reference samples were accurately assigned to either the North, Central, or East populations using leave-one-out self-assignment. All 35 reference samples from both the

North and East populations were assigned with 100% accuracy, and of the 35 birds from the Central population, 34 were correctly assigned.

Chinook salmon were accurately assigned to either the Sacramento or Klamath river basins even at read depths as low as 0.001X (Figure 2.6). All 12 test samples from the Sacramento river were correctly assigned at all read depth levels, and, of the 12 Klamath test fish, 11 were correctly assigned at all read depth levels, while one was correctly assigned at all read depth levels except for one of the five replicates at read depth 0.001X. The four samples with lowest full read depth (the four at the bottom of Figure 2.6) have log-likelihood ratios that are noticeably smaller than those of the remaining 20 fish at all downsampled read depth levels, possibly indicating that, in addition to being samples with low depth, they might also have yielded very poor quality DNA.

**Discussion**

Here, we present WGSassign and demonstrate its utility for population assignment with low-coverage WGS data. Our results, from both simulated and empirical data, show that low-coverage WGS data can be used to achieve high assignment accuracy even among weakly differentiated populations ($F_{ST} < 0.01$). We show that balancing effective sample size among populations is essential for avoiding assignment bias due to variation in the precision of allele frequency estimation for different populations. Effective sample size can also be used to guide decisions in study design for choosing the number of samples and sequencing depth in a given population. The ability to perform population assignment on large numbers of individuals, cost-effectively sequenced at low-coverage across the whole genome, further expands the utility of low-coverage WGS for population and conservation genomics.

*Performance of WGS*assign *and implications for population-assignment studies*

Our implementation of WGSassign allows users to perform population-assignment analyses from genotype likelihood data. Features of WGSassign include standard and leave-one-out (LOO) population assignment, as well as calculations of effective sample sizes (of both individuals and populations) and a *z*-score metric for determining whether an individual is from an unsampled population. Importantly, as implemented, these analyses can be parallelized across, which allows for fast computation of data produced from low-coverage WGS, even for computationally intensive applications such as LOO assignment. Studies of wild populations are typically limited in the number of samples available for sequencing, where 50 may be a large number of samples for a given population. With such a sample size, leave-one-out assignment at a standard low-coverage read depth of 1X could be expected to have a runtime on the order of for multiple populations and a million loci.

Implicit in standard population assignment tests is that there will always be a population with a maximum likelihood of assignment, even if the individual does not originate from any of the reference populations. To address this issue, we developed a *z*-score metric for testing whether an individual could be from an unsampled population. The *z*-score is based on the individual's observed likelihood of assignment in relation to the expected likelihood from a hypothetical individual from the same population with the same allele count data as the individual being tested. The *z*-score metric functions as expected at higher genetic differentiation ($F_{ST} > 0.05$) and with larger source populations by distinguishing the majority of individuals incorrectly assigned as having much lower *z*-scores (outside the 90% expected mass of the distribution of *z*-scores) than correctly assigned individuals. We recommend that any studies that

may have incomplete sampling coverage of all genetically distinct populations test for correct assignment with the *z*-score metric. However, since this metric is limited by sample size and genetic differentiation, a robust approach toward using it would involve, first, observing the metric's behavior by testing it upon individuals of known origin, calculating *z*-scores both for the population they are from and the other populations.

For high assignment accuracy, source populations need to have sufficient effective sample sizes in relation to genetic differentiation among the populations. However, individual samples being assigned can have extremely low read depth for accurate assignment. Our results from downsampled Chinook salmon data showed that individuals were still correctly assigned when individual samples had average read depths as low as 0.001X. This has powerful implications for population assignment studies, especially those that are conducted at a large scale. For example, in the mid-2000's an arduous, international, multi-laboratory study was undertaken to standardize a DNA database of 13 microsatellite loci for genetic stock identification of Chinook salmon at a coast-wide scale (Seeb *et al.* 2007). With today's sequencing power, a low-coverage WGS approach could provide a cost-effective method for creating a reference baseline of known populations without the need for extensive standardization of genetic makers. Fish of unknown origin could be sequenced at very low read depth, and still be accurately assigned to populations from the reference baseline.

A potential benefit of low-coverage WGS over other sequence data for population assignment, is that low-coverage WGS provides more markers for assignment to weakly differentiated populations. Population assignment studies with RADseq data have commonly used SNP filtering methods for selecting the most informative loci for assignment to weakly differentiated populations (DeSaix *et al.* 2019; Ruegg *et al.* 2014; Benestan *et al.* 2015). Further

identifying a subset of informative loci (e.g. < 200) can be cost-effective for genotyping large numbers of individuals for the purpose of assignment (Ruegg *et al.* 2014; Larison *et al.* 2021). However, our results highlight that high assignment accuracy is possible with low-coverage WGS data without the need for extensive analysis to determine the most informative loci. For example, high assignment accuracy was obtained with Yellow Warbler samples from weakly differentiated populations using 5,301,626 sites.

Furthermore, DNA quantity and quality requirements for RAD-seq methods—and even some chip-based genotyping methods—can be more stringent than they are for low-coverage whole genome sequencing. For example, reliable WGS data can be obtained from the tiny quantities of DNA adhering to the tip of a feather (Schweizer & DeSaix 2023), which is not possible with RAD-seq methods. Thus, being able to perform population assignment from low coverage whole genome sequencing data considerably expands the types of tissues available for sampling. And finally, using genotype data that is restricted to loci that are purposely biased toward detecting population structure (e.g. a SNP chip or hybridization-capture panel) limits the extent of analyses those data can be appropriately used for. Low-coverage WGS provides genome-wide data useful for population assignment in weakly differentiated populations, but it is also useful for demographic modeling, inference of population differentiation, detection of selection, and association studies (to name a few) because it has not been previously ascertained, and hence, biased.

*Accounting for population sample size and read depth with effective sample size*

Our development of the effective sample size metric provides a powerful tool for population genomics studies using low-coverage WGS data. Previous studies have provided

recommendations for the number of individuals and sequencing depth required to accurately estimate allele frequencies with low-coverage WGS data (Buerkle & Gompert 2013; Lou *et al.* 2021; Fumagalli 2013). Effective sample size provides a metric to quantify these recommendations and determine the precision of allele frequency estimation needed for different applications. For example, the recommendation of (Lou *et al.* 2021) of at least 10 individuals with 1X average sequencing depth for allele frequency estimation can be quantified as an effective sample size of 2.3 individuals in the simulations from this study (Figure 2.7). For assignment to populations with moderate to strong differentiation ($F_{ST} > 0.01$), population effective sample sizes of at least 2.3 individuals are sufficient for achieving consistently high assignment accuracy (Figure 2.3). However, at weaker genetic differentiation among populations, effective sample size needs to be increased for accurate assignment. Furthermore, for similar levels of effective sample size, populations with 10 samples tend to perform worse than populations with more samples. These results suggest that sequencing more individuals at lower read depths can be a more effective study-design strategy than sequencing fewer individuals at higher read depths. One reason that using more individuals for source populations may improve assignment accuracy is that it increases the likelihood of detecting low-frequency alleles.

Effective sample size can facilitate population-assignment study design by determining target numbers of individuals and average read depth for source populations. Our results show how effective sample size quantifies different study design options. For example, in our simulations a population with 10 samples with mean read depths of 1X had a mean effective sample size of 2.3 individuals. Increasing the total read depth of the population from 10X to 50X could be done by increasing the sequencing depth of the 10 individuals to 5X or increasing the

sampled number of individuals to 50 and keeping the mean individual sequencing depth at 1X. The simulation results show that increasing the sequencing depth produces an effective sample size of 7.2 individuals, while increasing sample size results in an effective sample size of 17.1 individuals (Figure 2.7). Quantifying the amount of information gain for different study designs can inform researchers on how to more efficiently allocate resources for sequencing efforts. Our simulation results show that disproportionate effective sample sizes among source populations can result in biased assignment of individuals to the populations with the highest effective sample sizes. We recommend that population assignment studies use the LOO assignment in WGSassign to determine if biased assignment is occurring. If all individuals across populations have similar average read depths, then subsetting source populations to the same number of samples for allele frequency calculation should remove this bias. However, different populations may tend to have higher or lower read depths, especially if different DNA sources are used, which will result in different effective sample sizes despite equal numbers of individuals. In this case, the individual effective sample size (Equation 13) output from WGSassign can be used to determine how many individuals to remove from the populations with the highest effective sample sizes. Alternatively, individuals could be further downsampled to reduce their effective sample size, which would decrease the overall population's effective sample size. Studies using low-coverage WGS data for population assignment can explore these different strategies with WGSassign to determine what is most effective for their datasets.

*Further improvements for population assignment*

Currently in our implementation of WGSassign, the issue of only a single allele being observed in a population, and thereby producing a likelihood of 0, is avoided by correcting a

population with a minor allele frequency of 0 by treating the locus as having a rare allele that would be observed in a single copy if another individual was to be sampled. Another approach that could potentially improve performance would be to specify a formal prior for the allele frequencies in each population (Rannala & Mountain 1997). Additionally, using a prior that accounts for the *a priori* expectation that allele frequencies at a locus are expected to be similar between weakly differentiated populations (Falush *et al.* 2003; Pella & Masuda 2006) may further improve performance of population assignment. We expect that the parameters of these more complex prior distributions could be estimated in an empirical Bayes approach (Maritz 2018) from the *n*-dimensional site frequency spectrum (Mas-Sandoval *et al.* 2022).

*Conclusion*

Low-coverage WGS is increasingly becoming more practical as sequencing costs decline and library preparation protocols are optimized for a wide-range of study systems (Schweizer & DeSaix 2023; Therkildsen & Palumbi 2017). In this paper, we present the WGSassign software which expands the types of analyses that can be done from genotype likelihoods. We demonstrate with simulated and empirical data that highly accurate and computationally efficient population assignment can be performed, even with weakly differentiated populations. We provide the software as open-source to facilitate further improvements on our developments in the field of molecular ecology.

## Tables and figures

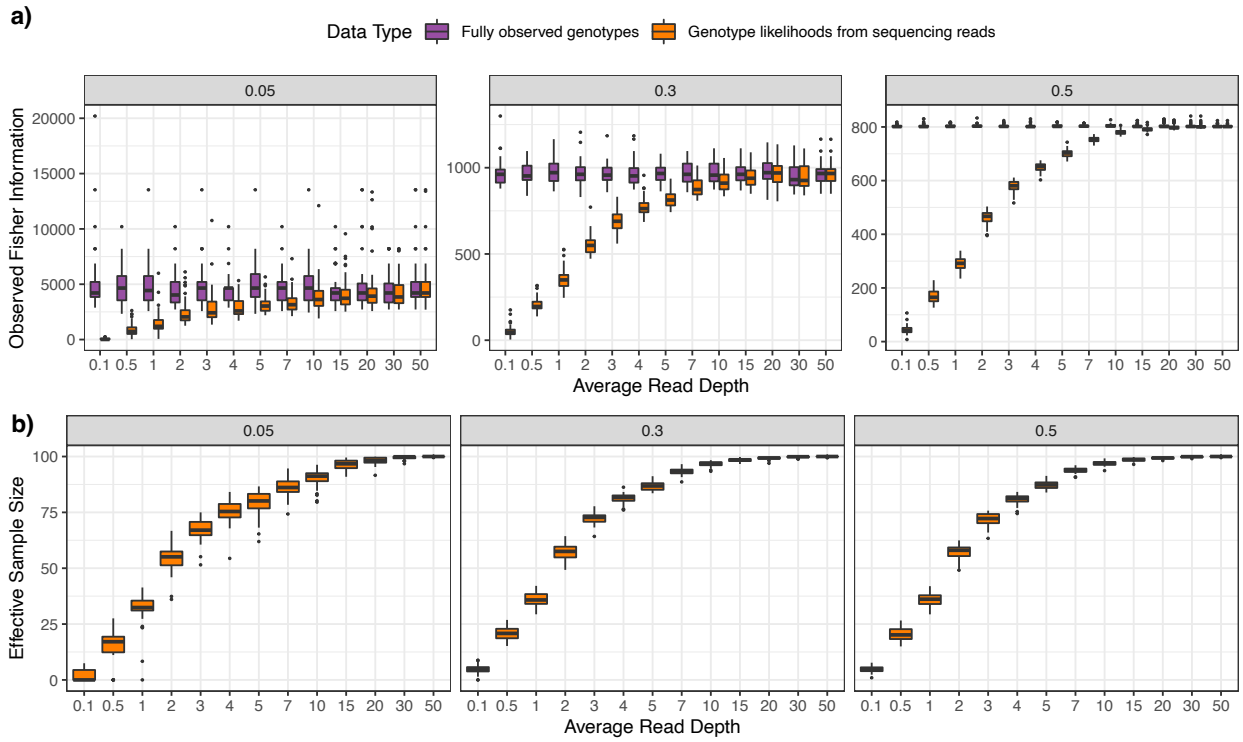


**Figure 2.1**. a) Observed information calculated for simulated data summarized either as fully observed genotypes (purple) or as genotype likelihoods (orange) computed from sequencing read data of different depths simulated from the genotypes. Fully observed genotype data is not affected by read depth, but an independent set of fully observed genotypes was simulated for each different value of read depth, and these are all shown in the figure. b) Effective sample sizes calculated for simulated genotype likelihood data. In each figure the facet headers give the true population allele frequency, the x-axis gives the average read depth in the simulations, and the distribution of quantities in the $y$ direction are summarized as boxplots showing the median (dark line) the first and third quartiles (the edges of the boxes) the largest (or smallest) value no further than 1.5x the interquartile range from the first and third quartiles (the whiskers) and outliers beyond the whiskers (individual points).

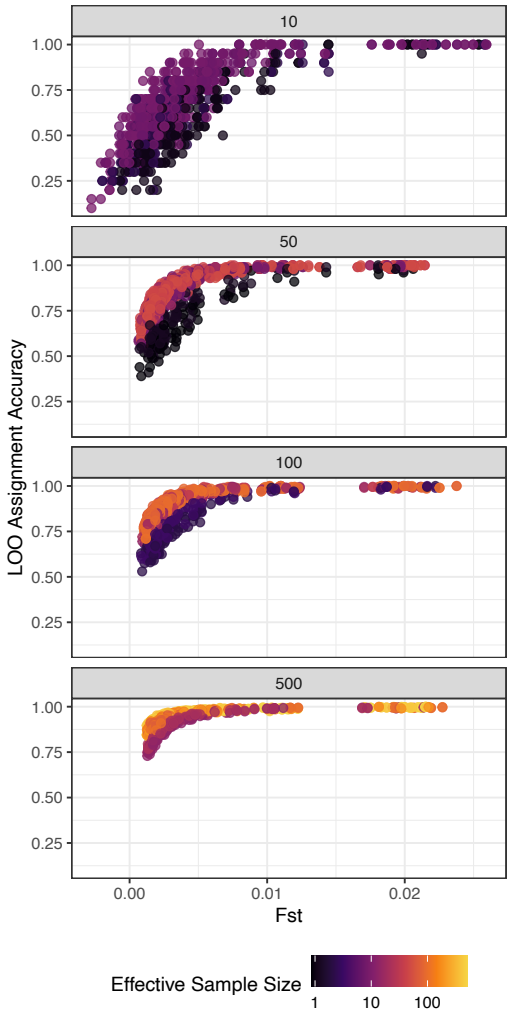


**Figure 2.2.** Leave-one-out (LOO) assignment accuracy for known source individuals increases as genetic differentiation increases. Each point represents a single one of 4,633 simulation runs of the two-population island model when effective sample sizes were greater than 0.1 individuals. Panels are ordered by the number of individuals (10, 50, 100, 500) sampled from each of the two populations. The proportion of correctly assigned individuals, via LOO cross-validation for one population is given on the y-axis and genetic differentiation between the two populations is on the x-axis. The points are colored by effective sample size log10 scale) of the population. Assignment accuracy in simulation runs with similar genetic differentiation tends to be greater for populations with greater effective sample size (lighter colors) than smaller effective sample sizes (darker colors). The variation in assignment accuracy decreases as more samples are used in the source population, with the highest amount of variation when 10 samples are used and the least amount of variation when 500 samples are used.

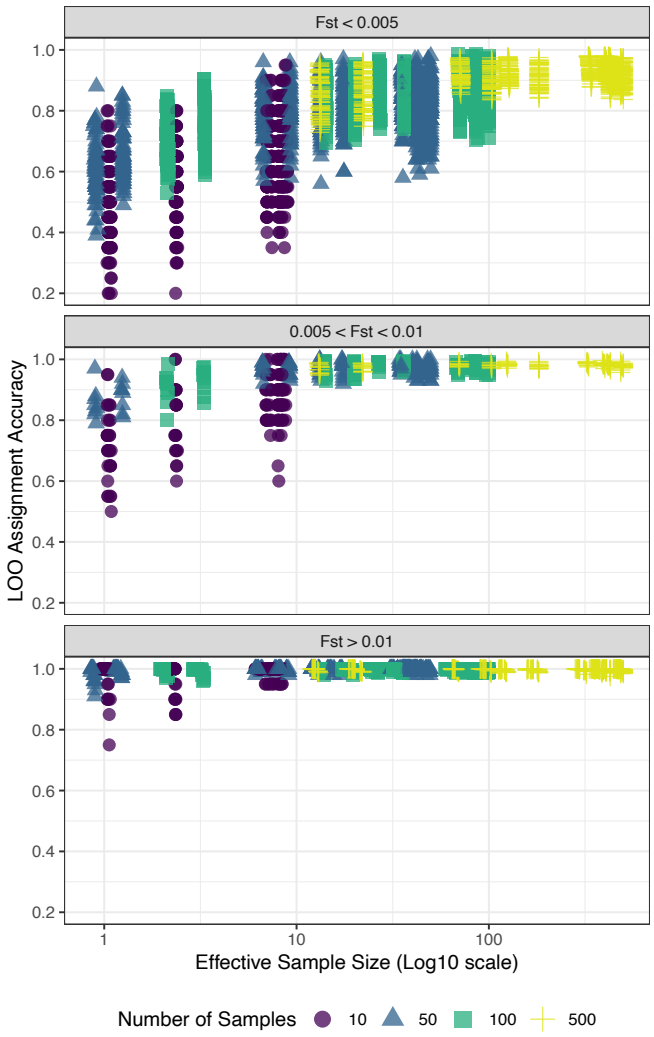


**Figure 2.3**. Increasing effective sample size results in an increase in LOO assignment accuracy. The proportion of correctly assigned individuals, using LOO cross-validation, for one population, is given on the y-axis and effective sample size (log10 scale) of the population is on the x-axis. Similar values of effective sample size results in a similar range of assignment accuracy, however the number of samples also influences the accuracy at lower effective samples sizes and with weaker genetic differentiation. Some of the effect of sample size, separate from effective sample size, can be explained by LOO assignment removing an individual from the source population during assignment, which will disproportionately decrease the precision of allele frequency estimation for smaller sample sizes than larger sample sizes.

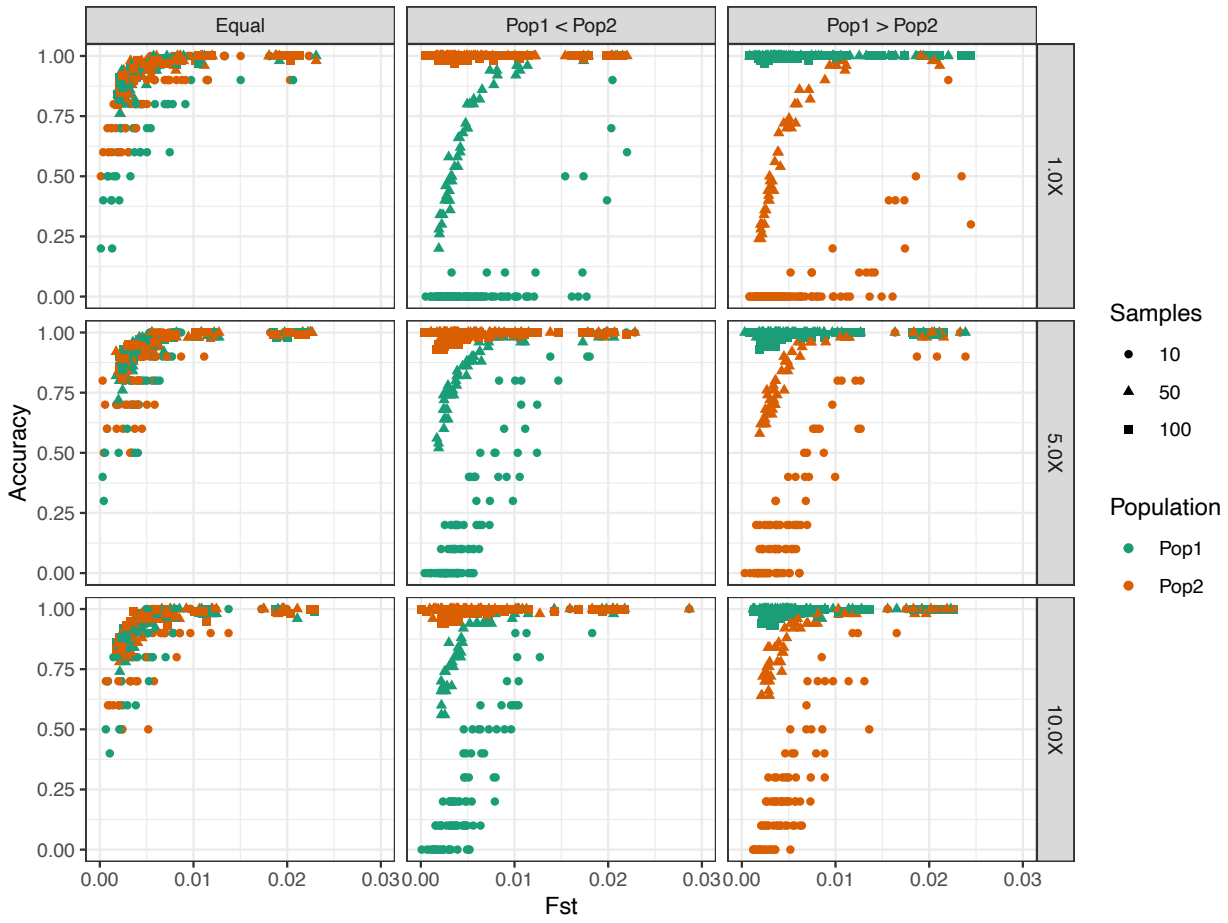


**Figure 2.4.** Unequal sample sizes among source populations result in decreased assignment accuracy due to differences in the precision of allele frequency estimation among the populations. Here, the two populations had either 10, 50, or 100 samples used for estimating allele frequency and then assigned via leave-one-out. When both populations had the same number of samples ("Equal" column), assignment accuracy generally increased as Fst increased and was similar for either population. When Population 1 had fewer samples than Population 2 ("Pop1 < Pop2" column), the assignment accuracy of Population 1 was generally less than that of Population 2, and the reverse was demonstrated when Population 1 had more samples than Population 2 ("Pop1 > Pop2" column). The reduction in assignment accuracy from biased sample sizes was also more pronounced with lower read depth.

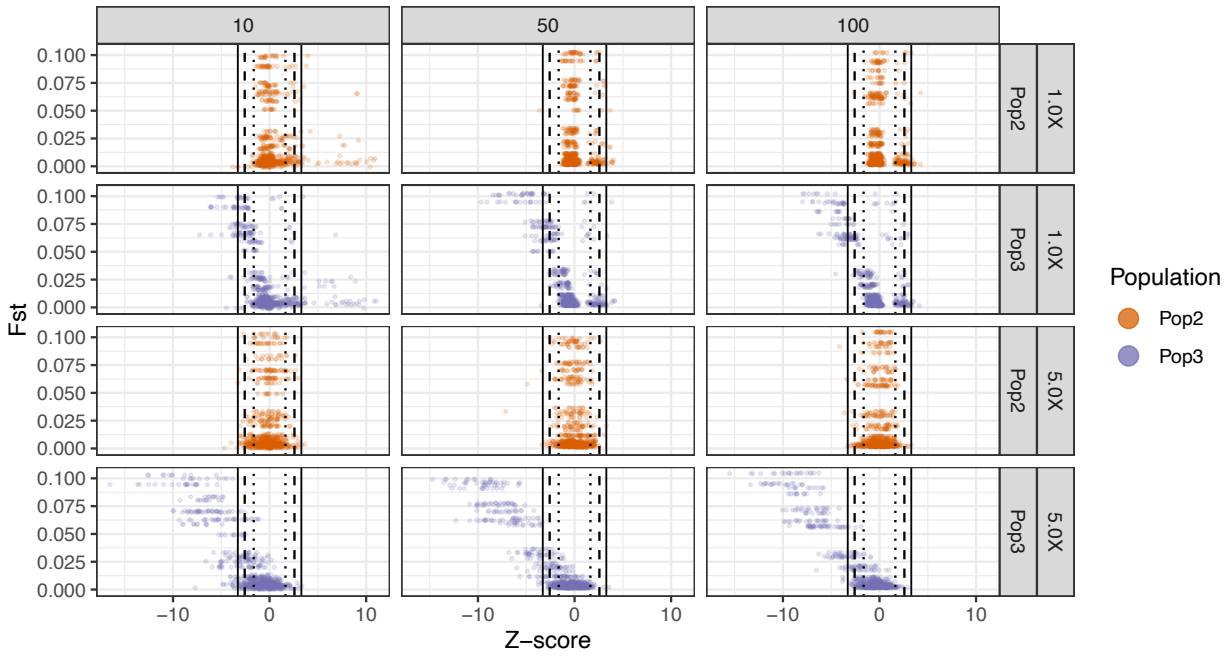


**Figure 2.5.** Results from the three-population stepping-stone model demonstrate the behavior of the z-score metric in identifying individuals from an unsampled population (Pop3) assigned to a population in the reference compared to individuals correctly assigned to their source population of origin (Pop2). Symmetric lines subtending 90%, 99%, and 99.9% of the mass of a standard unit normal random variate are given by vertical lines (dotted, dashed, and solid, respectively).

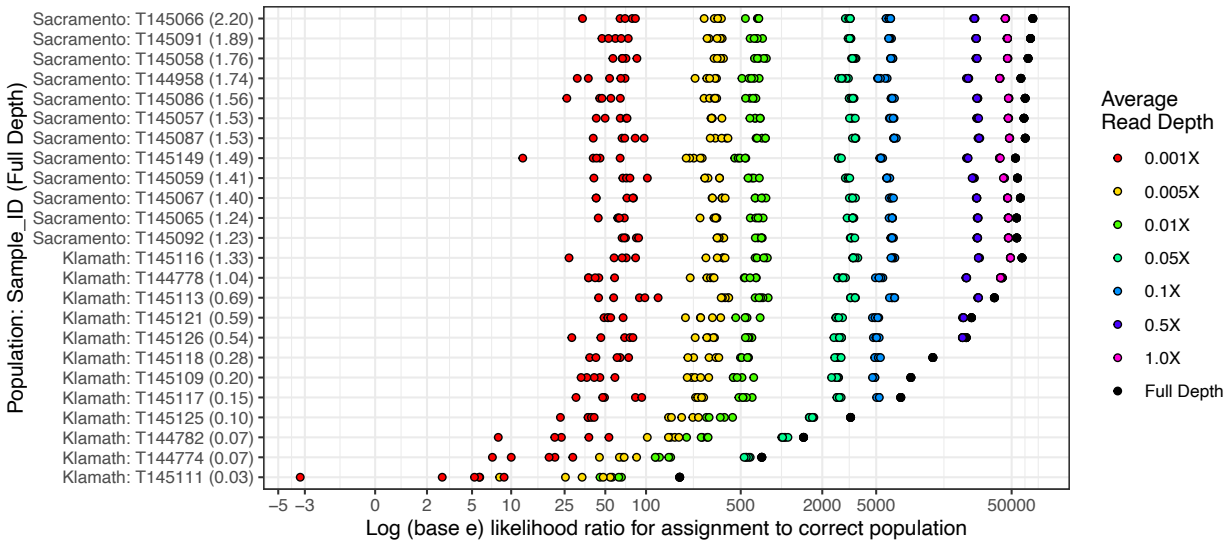


**Figure 2.6.** Log likelihood ratios for assignment at different read depth levels for the Chinook salmon data. On the $y$-axis are different Chinook salmon samples, labeled by their population, a colon, their ID number, and then in parentheses the average read depth of their aligned data at full depth. On the $x$-axis is the log-likelihood ratio in favor of assignment to their own (correct) population on a ``pseudo-log'' scale that accommodates negative values. Positive numbers indicate correct assignment. Colors denote the read depths after downsampling. There are five points for each individual at each value of downsampling, reflecting the 5 different seeds used for downsampling.

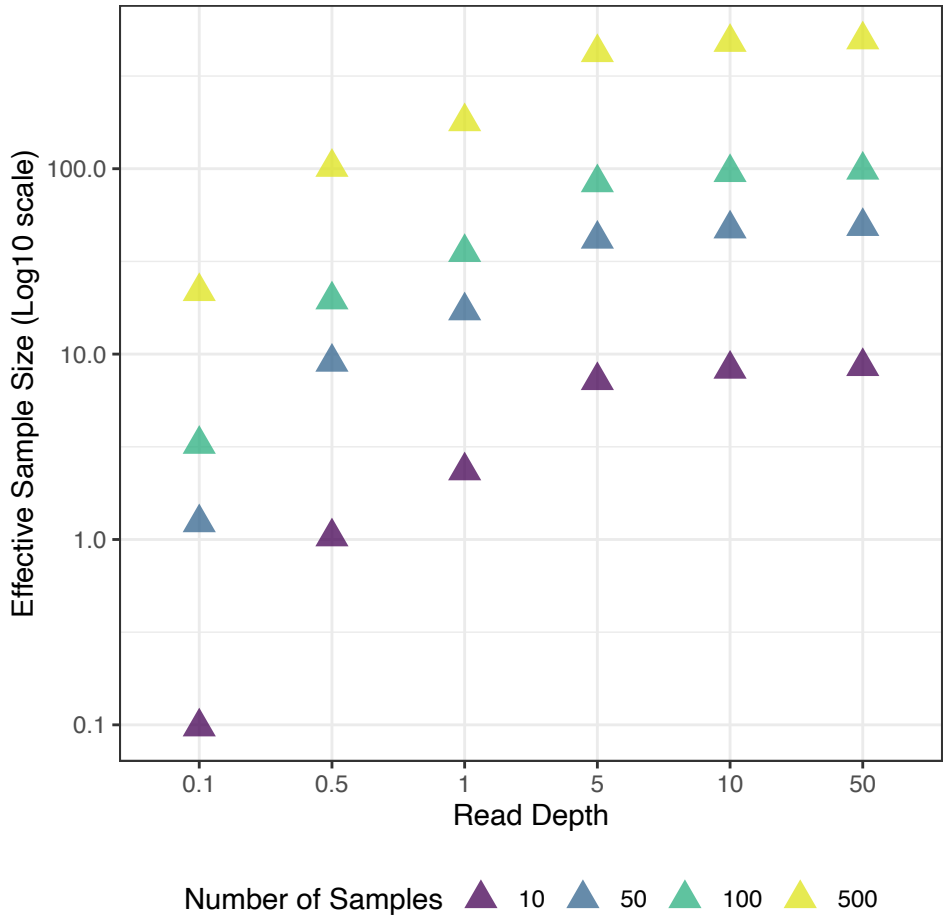


**Figure 2.7.** The relation between read depth and number of samples in determining the effective sample size highlights the potential for different sampling design strategies for achieving similar effective sample size. For example, if the target effective sample size is 10, then sequencing 500 individuals at 0.1x would likely overshoot the target, 50 individuals at 0.5x would be close to the target, and 10 individuals at >10x coverage would be close to the target.

LITERATURE CITED


Bay RA, Karp DS, Saracco JF, Anderegg WR, Frishkoff LO, Wiedenfeld D, Smith TB, Ruegg K (2021) Genetic variation reveals individual-level climate tracking across the annual cycle of a migratory bird. *Ecology Letters*, **24**, 819–828.

Benestan L, Gosselin T, Perrier C, Sainte-Marie B, Rochette R, Bernatchez L (2015) RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American Lobster (Homarus americanus). *Molecular ecology*, **24**, 3299–3315.

Buerkle AC, Gompert Z (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular ecology*, **22**, 3028–3035.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics*, **153**, 1989–2000.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* (2021) Twelve years of SAMtools and BCFtools. *Gigascience*, **10**, giab008.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

DeSaix MG, Bulluck LP, Eckert AJ, Viverette CB, Boves TJ, Reese JA, Tonra CM, Dyer RJ (2019) Population assignment reveals low migratory connectivity in a weakly structured songbird. *Molecular Ecology*, **28**, 2122–2135.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, **164**, 1567–1587.

Fisher RA (1922) On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, **222**, 309–368.

Fumagalli M (2013) Assessing the effect of sequencing depth and sample size in population genetics inferences. *PloS one*, **8**, e79667.

Fumagalli M, Vieira FG, Korneliussen TS, Linderoth T, Huerta-Sánchez E, Albrechtsen A, Nielsen R (2013) Quantifying population genetic differentiation from next-generation sequencing data. *Genetics*, **195**, 979–992.

Fumagalli M, Vieira FG, Linderoth T, Nielsen R (2014) ngsTools: methods for population genetics analyses from next-generation sequencing data. *Bioinformatics*, **30**, 1486–1487.

Gibbs HL, Dawson RJ, Hobson KA (2000) Limited differentiation in microsatellite DNA variation among northern populations of the yellow warbler: evidence for male-biased gene flow? *Molecular Ecology*, **9**, 2137–2147.

Kelleher J, Etheridge AM, McVean G (2016) Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, **12**, e1004842.

Kim SY, Lohmueller KE, Albrechtsen A, Li Y, Korneliussen T, Tian G, Grarup N, Jiang T, Andersen G, Witte D, *et al.* (2011) Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC bioinformatics*, **12**, 1–16.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: analysis of next generation sequencing data. *BMC bioinformatics*, **15**, 1–13.
Korneliussen TS, Moltke I (2015) NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, **31**, 4009–4011.
Larison B, Lindsay AR, Bossu C, Sorenson MD, Kaplan JD, Evers DC, Paruk J, DaCosta JM, Smith TB, Ruegg K (2021) Leveraging genomics to understand threats to migratory birds. *Evolutionary applications*, **14**, 1646–1658.
Larribe F, Fearnhead P (2011) On composite likelihoods in statistical genetics. *Statistica Sinica*, 43–69.
Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987–2993.
Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The sequence alignment/map format and SAMtools. *bioinformatics*, **25**, 2078–2079.
Lou RN, Jacobs A, Wilder AP, Therkildsen NO (2021) A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, **30**, 5966–5993.
Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods: matching biological questions with appropriate techniques. *Trends in Ecology & Evolution*, **20**, 136–142.
Maritz JS (2018) *Empirical Bayes methods with applications*, CRC Press.
Mas-Sandoval A, Pope NS, Nielsen KN, Altinkaya I, Fumagalli M, Korneliussen TS (2022) Fast and accurate estimation of multidimensional site frequency spectra from low-coverage high-throughput sequencing data. *GigaScience*, **11**.
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, **20**, 1297–1303.
Meisner J, Albrechtsen A (2018) Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, **210**, 719–731.
Nielsen R, Korneliussen T, Albrechtsen A, Li Y, Wang J (2012) SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. *PLoS ONE*.
Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, **12**, 443–451.
Paetkau D, Calvert W, Stirling I, Strobeck C (1995) Microsatellite analysis of population structure in Canadian polar bears. *Molecular Ecology*, **4**, 347–354.
Pella J, Masuda M (2006) The gibbs and split merge sampler for population mixture analysis from genetic data with incomplete baselines. *Canadian Journal of Fisheries and Aquatic Sciences*, **63**, 576–596.
R Core Team (2022) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences*, **94**, 9197–9201.
Ruegg KC, Anderson EC, Paxton KL, Apkenas V, Lao S, Siegel RB, DeSante DF, Moore F, Smith TB (2014) Mapping migration in a songbird using high-resolution genetic markers. *Molecular Ecology*, **23**, 5726–5739.

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (2022) GenBank. *Nucleic acids research*, **50**, D161.
Schweizer TM, DeSaix MG (2023) Cost-effective library preparation for whole genome sequencing with feather DNA. *Conservation Genetics Resources*, 1–8.
Seeb L, Antonovich A, Banks MA, Beacham T, Bellinger M, Blankenship S, Campbell M, Decovich N, Garza J, Guthrie Iii C, *et al.* (2007) Development of a standardized DNA database for Chinook salmon. *Fisheries*, **32**, 540–552.
Skotte L, Korneliussen TS, Albrechtsen A (2013) Estimating individual admixture proportions from next generation sequencing data. *Genetics*, **195**, 693–702.
Smouse PE, Waples RS, Tworek JA (1990) A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences*, **47**, 620–634.
Therkildsen NO, Palumbi SR (2017) Practical low-coverage genome wide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. *Molecular ecology resources*, **17**, 194–208.
Thompson NF, Anderson EC, Clemento AJ, Campbell MA, Pearse DE, Hearsey JW, Kinziger AP, Garza JC (2020) A complex phenotype in salmon controlled by a simple change in migratory timing. *Science*, **370**, 609–613.

# 3. LOW-COVERAGE WHOLE GENOME SEQUENCING FOR HIGHLY ACCURATE POPULATION ASSIGNMENT: MAPPING MIGRATORY CONNECTIVITY IN THE AMERICAN REDSTART (*SETOPHAGA RUTICILLA*)


## Summary

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

## Introduction

Long-distance migratory species pose distinct challenges to studies of ecology, evolution, and conservation because they occupy different geographical regions throughout the year that can be separated by thousands of kilometers. At each stage in the migratory annual cycle, migrant populations are subject to various stressors that can influence their fitness (Marra et al., 1998; Sillett et al., 2000). As a result, effective conservation efforts require understanding migratory connectivity, defined as the links between different geographic regions used across the annual cycle (Marra et al., 2015; Webster et al., 2002). In the past 20 years, population genetics has become a well-established means for tracking migratory populations, especially for studies involving large sample sizes or small-bodied individuals (Faaborg et al., 2010). However, the value of genetic markers is often limited by the amount of genetic differentiation in a species and the availability of genetic data from individuals across the annual cycle (Faaborg et al., 2010; Lovette et al., 2004).

Population assignment methods originated in the early 1980s and 1990s as a means of identifying breeding origins of migratory individuals back to distinct tributaries (in the case of fish) or geographic regions (in the case of bears) (Grant et al., 1980; Paetkau et al., 1995; Rannala & Mountain, 1997). Early methods relied on genetic markers that were limited to identifying only deep phylogeographic breaks within species (Kimura et al., 2002; Lovette et al., 2004; Ruegg & Smith, 2002). In recent years, next generation sequencing has facilitated the screening of a significantly larger number of genetic markers allowing for the delineation of breeding populations at finer spatial scales (Battey et al., 2018; DeSaix et al., 2019; Ruegg et al., 2014). Cost-effective delineation of patterns of migratory connectivity was made possible by designing single nucleotide polymorphisms (SNP) assays for a subset of these markers that were

particularly useful for population assignment (Larison et al., 2021; Rueda-Hernández et al., 2023; Ruegg et al., 2014). While recent reductions in the cost of whole genome sequencing have made it possible to directly use low-coverage whole genome sequencing (lcWGS) data to screen migrant samples, the lack of software capable of dealing with the increase in marker number has prevented this method from being used for population assignment (DeSaix et al. in review).

Low-coverage WGS has made sequencing more affordable for non-model organisms by reducing the sequencing effort per individual, however it has distinct challenges. One of these challenges is dealing with low sequencing read depths per individual, which necessitates the use of probabilistic frameworks for genotype calling to account for the uncertainty inherent in the data (Nielsen et al., 2011, 2012). Accurate estimates of parameters such as allele frequency can be obtained by prioritizing larger sample sizes of individuals with lower sequencing depth (Buerkle & Gompert, 2013; Fumagalli et al., 2013; Lou et al., 2021). Guidelines for achieving accurate allele frequency estimation with lcWGS include sequencing individuals at a minimum of 1X coverage (Buerkle & Gompert, 2013) or having at least 10 individuals sequenced with a total sequencing depth of at least 10X (Lou et al., 2021). To take advantage of lcWGS data for population assignment, DeSaix et al. (*in review*) recently developed a software package, WGSassign, that accounts for uncertainty inherent to lcWGS data in population assignment tests. Results from extensive simulations, as well as two empirical data sets, demonstrated that accurate assignment with lcWGS data is possible for weakly differentiated populations (DeSaix et al., *in review*). Here, for the first time, we use lcWGS data to assign migrants to their population of origin.

The American Redstart (*Setophaga ruticilla*) is an ideal system for evaluating the potential gains in effectiveness achievable by using lcWGS data for population assignment

because previous studies using a variety of methods provide a strong foundation for comparisons. The American Redstart is a widely distributed migratory songbird with a breeding distribution across North America and stationary nonbreeding distribution throughout the Caribbean, northern South America, Central America, and Mexico (Sherry et al., 2016). For several decades, the American Redstart has been a model species for understanding migratory ecology and has been used to elucidate territoriality on the wintering grounds (Marra et al., 1993), foraging behavior (Lovette & Holmes, 1995), habitat selection (Marra & Holmes, 2001; Sherry & Holmes, 1996), and carry-over effects of stressors across the annual cycle (Marra et al., 1998; Studds & Marra, 2011). Phylogeographic structure has previously been detected between a small region in the Maritime Provinces, specifically in Newfoundland and New Brunswick in the northeastern portion of the range, and the rest of the continental breeding range using mtDNA (Colbeck et al., 2008). Subsequent analysis of migratory connectivity (i.e., the migratory connections between breeding and nonbreeding habitats across a species' annual cycle) using mtDNA revealed that Newfoundland breeders overwintered on the islands of Puerto Rico and the Dominican Republic, while continental breeding birds overwintered across the entire nonbreeding range (DeSaix et al., 2022). Stable isotope studies have shown strong migratory connectivity, with eastern breeding birds overwintering in the Caribbean and western breeding birds overwintering in Central America and Mexico (Marra et al., 1998; Norris et al., 2006; Studds et al., 2021), but whether these migratory differences correspond to genetic differentiation has not been tested.

Here we aim to demonstrate the effectiveness of using lcWGS data for population assignment of nonbreeding individual using the American Redstart as a model species. Our main objectives were: 1) Identify population-specific migratory connectivity in the American Redstart

using lcWGS data, 2) Assess conservation implications of migratory connectivity by identifying relative abundance and trends in population size, and 3) Provide study design recommendations to facilitate the use of lcWGS data in other population assignment studies. Our results have broad implications for improving our understanding of the ecology and evolution of migratory species through conservation genomics approaches.

## Methods

### *Genetic sampling and library preparation*

Sample site locations were chosen to maximize sampling coverage across the breeding and nonbreeding ranges of the American Redstart. We used genetic samples from a total of 330 individuals: 182 individuals from 16 locations across the breeding range and 148 individuals from 15 locations in the nonbreeding range. Sample collection occurred between 1993 and 2022 and consisted of either blood from brachial venipuncture or feathers. We extracted DNA from blood samples using the standard protocol for Qiagen DNEasy Blood and Tissue Kits and we modified the protocol to maximize DNA yield from feathers (Schweizer & DeSaix, 2023). Whole genome sequencing libraries were prepared following modifications of Illumina's Nextera Library Preparation protocol (Schweizer & DeSaix, 2023). Pooled libraries were sequenced on eight HiSeq 4000 lanes at Novogene Corporation Inc with a target sequencing depth of 2X per individual.

### *Bioinformatics*

We trimmed the sequence data to remove potential PCR artifacts using the program TrimGalore version 0.6.5 (https://github.com/FelixKrueger/TrimGalore), a wrapper for Cutadapt

(Martin, 2011). We used the Burrows-Wheeler Aligner software version 0.7.17 (Li & Durbin, 2009) to map reads to a reference genome from the closely related Yellow Warbler (*Setophaga petechia*; Bay et al. 2018). After mapping, the resulting SAM files were sorted, converted to BAM files, and indexed using Samtools version 1.9 (Li et al., 2009). We marked read duplicates with MarkDuplicates from GATK version 4.1.4.0 (McKenna et al., 2010) and clipped overlapping reads with the clipOverlap function from bamUtil (https://genome.sph.umich.edu/wiki/BamUtil:_clipOverlap). Sequencing depth for individuals was calculated using MEAN_COVERAGE output from the CollectWgsMetrics function from GATK (McKenna et al., 2010) which specifies genomic coverage excluding reads that do not pass quality filters. Initial population genetics analyses revealed a large effect in the data due to high variation in sequencing depth among individuals. To reduce sequencing depth variation, we followed the recommendations of Lou & Therkildsen (2022) and used the DownsampleSam function from GATK to randomly down sample reads from BAM files with greater than 2X coverage, to 2X coverage.

To identify genetic markers from low-coverage WGS data, we used stringent filtering options in ANGSD version 0.9.40 (Korneliussen et al., 2014). We retained reads with a mapping quality of at least 30 and base quality of at least 33. SNPs were identified based on a p-value of less than 1e-6. We retained SNPs that had read data in at least 50% of individuals ($n$ = 165), a minor allele frequency greater than 0.05, and minimum and maximum total depths of 231 and 924, respectively. The minimum total depth threshold was chosen by the minimum number of individuals required to call a variant ($n$ = 165) multiplied by the mean sequencing depth of all individuals (1.4X). The maximum total depth threshold was determined by 2 * total number of

individuals * mean sequencing depth. The filtered variants were output as genotype likelihoods and used in subsequent analyses.

*Genetically distinct breeding populations*

Given that signatures of population structure can be skewed by closely related individuals, we used NGSrelate version 2 (Hanghøj et al., 2019; Korneliussen & Moltke, 2015) to identify and remove individuals with with up to second-degree relationships (kinship > 0.0884). We implemented principal components analysis (PCA) and estimated individual admixture proportions using Pcangsd (Meisner & Albrechtsen, 2018), which estimates individual allele frequencies to minimize bias from low and variable sequencing depth. We determined the number of genetically distinct breeding populations of American Redstarts by identifying congruent geographic signatures of clustering in the PCA with groupings of individuals based on admixture proportions. Posterior probabilities of group membership from the admixture proportions were visualized on a base map from Natural Earth (naturalearthdata.com) with each group specified by a different colour, and clipped to the breeding range of the American Redstart (Strimas-Mackey et al., 2021). Colour transparency was scaled such that the highest posterior probability of group membership is opaque while the smallest posterior probability is transparent. Visualization was performed in R (R Core Team, 2021). Genetic differentiation among the breeding populations was calculated by creating site allele frequency files for each breeding population and calculating $F_{ST}$ in ANGSD (Korneliussen et al., 2014). In addition to summarizing global $F_{ST}$ values for pairwise population genetic differentiation, we performed sliding window calculations of $F_{ST}$ across the genome using 50kb windows and 10kb steps.

*Effective sample size and population assignment*

To assess our ability to accurately assign individuals of unknown origin to breeding populations, we first determined the accuracy of assignment of the known breeding origin individuals using a leave-one-out approach implemented in WGSassign (DeSaix et al. *in review*). Leave-one-out avoids assignment bias by iteratively removing an individual from their given source population, re-estimating the allele frequency of the source population, and then calculating the likelihood of the individual's assignment to each population. Otherwise, assignment accuracy results will be upwardly biased due to individuals' genotype information being present in the allele frequency estimates of the source population they are being assigned to. Another source of bias in assignment tests is variation in the precision of allele frequency estimation, which arises from populations having different numbers of samples and/or having differences in sequencing depth of their individuals. To mitigate this bias, we tested two other approaches for source population sampling design: 1) we randomly subsampled individuals from breeding so all breeding populations had the same number of individuals as the population with the fewest samples (size-standardized breeding populations; SSBPs) and 2) we followed the guidelines in DeSaix et al. (*in review*) to subsample individuals from the breeding populations to standardize the *effective sample sizes* of the breeding populations (effective-size-standardized breeding populations; ESSBPs). *Effective sample size* is a Fisher information metric that determines the comparable number of individuals with known genotypes that would reflect the same variance in estimated allele frequency as the sampled low-coverage individuals (DeSaix et al., *in review*). The purpose of ESSBPs is to equalize the effective sample size among populations by removing individuals from the populations with the highest effect sample sizes, thereby making the precision of allele frequency estimation similar among the different

populations. We used WGSassign to calculate each breeding population's effective sample size for the SSBPs and ESSBPs and performed leave-one-out assignment. We also performed standard assignment with all breeding individuals that had been removed from the SSBPs and ESSBPs. Leave-one-out assignment for the full data set and the combined leave-one-out assignment and standard assignment accuracy were compared across all three source population sampling designs. Posterior probabilities of assignment to a population were determined by dividing the maximum likelihood of assignment over the sum of all likelihoods. A cut-off of 0.8 was used for the posterior probability to determine if an individual was confidently assigned to a population.

*Low-coverage and population assignment*

Since the majority of our nonbreeding and breeding samples were feathers and blood, respectively, we expected the nonbreeding samples to have lower sequencing depth than breeding samples. Therefore, to ensure that we could still achieve high assignment accuracy at lower depths for the nonbreeding samples, which have unknown breeding origin, we first tested assignment accuracy with low coverage breeding samples of known origin. We used the set of individuals from our ESSBPs to estimate population allele frequencies (our training set) and used the remaining breeding samples as a test set. We created two data sets from the test set individuals by randomly down sampling reads from the BAM files of these individuals to 0.1X and 0.01X using the DownsampleSam function from GATK (Mckenna et al., 2010). These two thresholds were based on the majority of the nonbreeding samples being greater than 0.1X and the lowest coverage sample being 0.02X. To determine the accuracy of assignment of individuals

with low sequencing depths, we assigned the test sets back to the standardized breeding populations and compared the population assigned with the known population of origin.

*Determining breeding origin of individuals on the nonbreeding range*

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

*Demographic analysis*

We estimated relative population size indices and population trends (1968-2021) for each of the five breeding populations and across the entire breeding range using Breeding Bird Survey (BBS) data (Pardieck et al., 2020). The BBS provides standardized detection data of avian species during the main part of the breeding season (June) which is collected by observers along

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

## Results

*Genetically distinct breeding populations*

Sequencing efforts resulted in sequences from 330 individuals with a mean coverage of 1.6X (range: 0.02X – 5.2X). For the 182 breeding samples, the mean coverage was 1.7X (range:

0.6X – 5.2X), while the 148 nonbreeding samples had a mean coverage of 1.5X (range: 0.02 – 3.4X). Down sampling individuals above 2X coverage to 2X coverage, resulted in an overall mean coverage of 1.4X (1.5X for breeding samples, 1.3X for nonbreeding samples). Our SNP filtering produced genotype likelihood data for 4,722,390 variants. We removed 13 individuals from subsequent analyses due to high relatedness by removing a single individual from each related pair. Principal components analysis with the breeding samples revealed five genetic clusters that aligned with geography: Western Boreal (Alaska to Saskatchewan), Basin Rockies (South Dakota and Montana), Southern Temperate (from Missouri, east to Maryland, south to Louisiana), Northern Temperate (from Minnesota, east to Quebec, south to Pennsylvania), and Maritime Provinces (New Brunswick and Newfoundland). Admixture results for five groups revealed a similar delineation of individuals as in the principal components analysis (Figure 3.1). Pairwise $F_{ST}$ values among these breeding populations had a mean of 0.009 and ranged from the weakest differentiation ($F_{ST}$ = 0.004) between the Northern Temperate and Southern Temperate groups and the strongest ($F_{ST}$ = 0.018) between Maritime Provinces and Basin Rockies. Genome-wide analysis of $F_{ST}$ revealed that the generally weak genetic differentiation among populations was punctuated by regions of elevated genetic differentiation. Based on our comprehensive sampling of the core regions of the American Redstart breeding range and the population structure results, we expect there to be no unsampled genetically distinct breeding populations. For example, while we did not sample individuals from Alaska, we expect Alaska breeding birds to be a part of the Western Boreal breeding population.

*Effective sample size and assignment*

Breeding populations ranged in number of samples from 27 (Maritime Provinces and Basin Rockies) to 47 (Southern Temperate) and ranged in effective sample size from 12.3 – 24.6. Mean accuracy of leave-one-out assignment with these individuals was 89.3% (151 out of 169 individuals), and accuracy by breeding population ranged from 63.0% (Maritime Provinces) to 100% (Southern Temperate and Basin Rockies). All 18 individuals that were inaccurately assigned were assigned to a breeding population with higher effective sample size than the known breeding population (Figure 3.2). All posterior probabilities of assignment (accurate and inaccurate) were greater than 0.8.

The SSBPs all had 27 samples but ranged in effective sample size from 12.3 – 16.1. The mean accuracy of assignment was 97.0% (164 out of 169 individuals) for the leave-one-out assignment with the 135 individuals in the SSBPs and standard assignment for the remaining 34 individuals. Three individuals were incorrectly assigned in the leave-one-out assignment test and two individuals were incorrectly assigned in the standard assignment test. All five incorrectly assigned individuals were assigned to a breeding population with higher effective sample size than the known breeding population (Figure 3.2).

The ESSBPs ranged in number of samples from 21 (Basin Rockies) to 27 (Maritime Provinces) but had minimal variation in their effective sample sizes (range: 12.0 – 12.5). Mean assignment accuracy was 99.4% (168 out of 169 individuals) for the leave-one-out-assignment with the 122 individuals in the ESSBPs and standard assignment for the remaining 47 individuals. Only one individual was incorrectly assigned from the Northern Temperate population to the Southern Temperate population, and this did not correspond to a breeding population with higher effective sample size. Interestingly, this same individual (sampled in

Minnesota from the Northern Temperate population) also stands out in the PCA results as clustering more closely with individuals from the Southern Temperate population. Given the higher accuracy of assignment with the ESSBPs, compared to the other sets of breeding individuals, we continued subsequent assignment testing using only the ESSBPs data set as the source populations.

*Testing sequence depth thresholds for assignment*

To test whether lower coverage data would affect our ability to accurately assign breeding individuals, we used the 47 individuals from the breeding populations, that were not used in the ESSBPs, as a testing set for further down sampling. We did not down sample individuals from the ESSBPs because we did not want to lower the effective sample size of the source populations, but rather test how well we could assign individuals with lower coverage given the effective sample sizes of our source populations (and the amount of genetic differentiation among them). The testing set consisted of 22 individuals from the Southern Temperate population, 14 individuals from the Northern Temperate population, 6 individuals from Basin Rockies population, and 5 individuals from the Western Boreal population. The sequencing depth of these 47 individuals ranged from 0.6X – 2.0X. Down sampling these individuals to 0.1X resulted in 100% assignment accuracy, and further down sampling to 0.01X resulted in 97.9% accuracy (one individual from the Southern Temperate population assigned to the Northern Temperate population; Table 3.1). The individual incorrectly assigned from the Southern Temperate population was from a sampling site in Pennsylvania which is on the border of our boundary for the Southern Temperate and Northern Temperate populations.

*Nonbreeding assignment*

Implicit in the assignment of the nonbreeding individuals to breeding populations is that the breeding origin of these individuals is unknown. Given that assignment to the ESSBPs had an accuracy of 99.4% (168 out of 169 samples) for individuals with sequencing coverage of 0.6X – 2.0X, and accuracy of 97.9% (46 out of 47 individuals) for individuals down sampled to sequencing coverage of 0.01X, we assumed that we could correctly assign nonbreeding individuals (sequencing coverage range: 0.02X – 2.0X, mean 1.3X) with high confidence. Assignment of the 148 nonbreeding individuals resulted in the largest number of individuals being assigned to the Northern Temperate population (n = 64) and the least number of individuals being assigned to the Basin Rockies population (n = 2). Of the 148 individuals, 139 individuals had assignment consistency of at least 0.8 for the 10 subsets of data, and these individuals were used to infer migratory connectivity. Testing consistency of assignment on the 47 breeding individuals identified three individuals with assignment consistency of < 0.8. One of these individuals from Minnesota was previously identified as an outlier in the PCA, and the other two individuals were from Pennsylvania, which is on the boundary of the Southern Temperate and Northern Temperate populations.

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

*Demographic analysis*

Using 1,766 BBS routes from 1968-2021, we estimated the range-wide trend in population size to be -0.29% per year (95% CI: -0.57, -0.02). Trends among the breeding populations were variable (Table 3.2). The Northern Temperate breeding population was estimated to be increasing by 0.67% per year (95% CI: 0.33, 1.01). The Southern Temperate breeding population was estimated to be declining by 0.34% per year (95% CI: -0.75, 0.06), but the credible intervals were overlapping 0, thus indicating potential stability in that population. The remaining three populations (Basin Rockies, Maritime Provinces, and Western Boreal) were all estimated to be declining and had negative values for the upper bounds of the 95% credible intervals. The Northern Temperate population had the highest relative abundance of 3.70 (95% CI: 3.09, 4.52), followed by Maritime Provinces (1.96; 95% CI: 1.57, 2.49), Western Boreal (0.66; 95% CI: 0.52, 0.84), Southern Temperate (0.15; 95% CI: 0.13, 0.17), and Basin Rockies (0.01; 95% CI: 0.01, 0.02). The density of the number of birds per BBS route was highest in the Maritime Provinces (30.56; 95% CI: 24.33, 38.60), followed by Northern Temperate (15.16; 95% CI: 12.68, 18.54), Western Boreal (1.66; 95% CI: 1.31, 2.10), Southern Temperate (0.67; 95% CI: 0.58, 0.78), and Basin Rockies (0.16; 95% CI: 0.10, 0.25).

**Discussion**

The results of this study demonstrate that lcWGS data is well-suited for highly accurate population assignment, even with weakly differentiated population structure. In the American Redstart, lcWGS data provided an improvement over previous migratory connectivity studies using genetic and stable isotope data (DeSaix et al., 2022; Norris et al., 2006; Studds et al., 2021) by allowing us to identify five genetically distinct breeding populations and clearly delineate

population-specific nonbreeding ranges. Identifying migratory connectivity of genetically distinct populations is an essential step toward full annual cycle conservation aimed at preserving unique genetic variation. To this end, we integrate the migratory connectivity results with analysis of population abundance and trends to demonstrate the conservation implications of the observed population-specific migratory patterns. More broadly, we also show that when using lcWGS data for population assignment it is essential to implement a sampling design that balances *effective sample size* across source populations to avoid assignment bias that arises from variation in sequencing depth and population sample size.

*Mapping migratory connectivity*

Population structure analyses identified five genetically distinct breeding populations with weak genetic differentiation, in contrast to a previous mtDNA analysis that identified only two populations split by a phylogeographic break (Colbeck et al., 2008). Our delineation of the Maritime Provinces breeding population in the far northeast portion of the range corresponds with the Newfoundland population from the mtDNA analysis. Colbeck et al. (2008) hypothesized that the phylogeographic separation of Newfoundland and mainland American Redstart populations was the result of two refugia during Pleistocene glaciations. Our findings of weak genetic differentiation between the Maritime Provinces and other breeding populations suggest that there is ongoing gene flow among these populations. However, the limited admixture of individuals sampled in Newfoundland supports the notion that geographic separation of the island provides some barrier to gene flow, which has been demonstrated in several other avian species (Ralston et al., 2021). The weakest genetic differentiation was found among the Western Boreal, Northern Temperate, and Southern Temperate breeding populations

($F_{ST}$: 0.004-0.006), suggesting limited barriers to gene flow. The Basin Rockies breeding population had higher genetic differentiation with the eastern breeding populations than the more northern Western Boreal population, which corresponds to the Great Plains functioning as a barrier to gene flow.

Using the five genetically distinct breeding populations allowed us to document at a fine scale more complex migratory patterns than previously identified. At the continental scale, our results broadly correspond to previous stable isotope studies that found eastern breeding American Redstarts overwintered in the eastern nonbreeding range and western breeders overwintered in the west (Norris et al., 2006; Studds et al., 2021). In several other species of Nearctic–Neotropical migrants, similar patterns of parallel migration have also been observed (Fraser et al., 2013; Garcia-Perez & Hobson, 2014; González-Prieto et al., 2017; Rushing et al., 2014). However, in contrast to previous isotope analyses in American Redstart (Norris et al., 2006; Studds et al., 2021), our use of genomic data allowed us to clearly differentiate Maritime Provinces and Northern Temperate breeding birds and revealed that individuals breeding in the Maritime Provinces do not follow the parallel migration pattern. Parallel migration would result in these breeders being found in the far eastern portion of the nonbreeding range (e.g., Lesser Antilles and Trinidad and Tobago). Instead, individuals from the Maritime provinces bypass the Caribbean portion of the nonbreeding range and have a "leap-frog" migratory pattern to eastern Colombia. One explanation for the discordance of the Maritime Provinces migratory connectivity patterns from the rest of the breeding populations is the phylogeographic separation of these regions documented by Colbeck et al. (2008). Migration routes are influenced by the historical separation of Pleistocene glacial refugia (Newton, 2008; Ruegg et al., 2006) and in the American Redstart, an Atlantic Shelf (near the Maritime Provinces) and eastern continental

refugia are hypothesized to have caused the observed phylogeographic separation of these regions (Colbeck et al., 2008). A previous mtDNA analysis of American Redstart migratory connectivity only detected several individuals from the Maritime Provinces population in the Caribbean islands of the Dominican Republic and Puerto Rico but lacked samples from South America (DeSaix et al., 2022). Our results suggest that the Maritime Provinces breeding population has the strongest connectivity with eastern Colombia, but given our limited sampling across South America the full extent of the population's connectivity across South America is unknown.

Our use of genomic data allowed us to characterize migratory connectivity at a fine scale and identify distinct regions on the wintering range that separate breeding populations. In southern Central America, a clear split between the two sampling sites in Costa Rica, a separation of 360 km, occurs where the northern site predominantly has individuals from the Western Boreal breeding population and the southern site has individuals from the Southern Temperate population. This split corresponds with a biogeographic separation of drier broadleaf forest in the northern Pacific side of Costa Rica and moist broadleaf forest in the southern Pacific side (Corrales, Bouroncle, & Zamora 2015). Another geographic split in breeding origin occurs between the Lesser Antilles (Southern Temperate) and the Greater Antilles (Northern Temperate) in the Caribbean. Sampling of American Redstarts in Colombia was limited to the eastern slopes of the East Andes, and may not represent the wider Andes, given that the three chains of the Andes that run through Colombia influence connectivity patterns in the Canada Warbler, *Cardellina canadensis* (González-Prieto et al., 2017). Further population assignment studies that include sampling of American Redstarts from the Central and Western Andes, and

the Caribbean region of Colombia, may identify the Andes Mountains as another barrier in the nonbreeding region, creating geographic splits in breeding origin for this species.

*Conservation implications*

Describing migratory connectivity is essential for informing effective wildlife conservation and management decisions involving migratory species (Martin et al., 2007; Small-Lorenz et al., 2013). Our delineation of five breeding populations, and their linkages to wintering regions, provides the necessary information to prioritize regions for conservation (Ruegg et al., 2020) and improve our understanding of the underlying drivers of abundance. While our demographic analysis highlights that the species is declining overall from 1966 – 2021, there is wide variation in trends among the breeding populations. The Northern Temperate population has the largest population of American Redstarts on the breeding grounds and is increasing in abundance. One potential explanation for the increase in abundance is that birds in the southern portion of the breeding range have shifted their breeding latitude northward in response to climate change, as has been documented in other Nearctic-Neotropical migrants (Gómez et al., 2021; Rushing et al., 2020). In this scenario, genetic differentiation would also likely erode between the Northern Temperate and Southern Temperate populations and our genetic differentiation results do highlight these two populations as having the weakest genetic differentiation ($F_{ST}$ = 0.004) of all breeding population comparisons. However, our demographic results do not depict a correspondingly large decline in the Southern Temperate breeding population which would be the source population of northward movement. Given that the Northern Temperate breeding population has strong connectivity with the Greater Antilles archipelago in the Caribbean, efforts aimed at conserving the greatest proportion of the global

distribution of American Redstarts could focus on the Northern Temperate-Greater Antilles migratory population.

The Maritime Provinces population had the second highest abundance of American Redstarts and the highest density of individuals. Despite being geographically adjacent to the Northern Temperate breeding population, Maritime Provinces individuals were detected almost exclusively outside the Caribbean, along the eastern slopes of the Andes of Colombia. Our demographic analysis highlighted the Maritime Provinces to be the second fastest declining population. Thus, future research into the stressors driving this decline could focus on the breeding region as well as stationary nonbreeding region of eastern Colombia. Notably, other populations of long-distance migratory birds connected to the Eastern Andes are also experiencing declines, including populations of Canada Warbler (Wilson et al., 2018) and Cerulean Warbler, *Setophaga cerulea* (Raybuck et al., 2022). Additionally, in species such as the Canada Warbler, migration routes between North and South America can concentrate in small regions of Central America which can also affect population trends (Roberto-Charron et al., 2020). Given the phylogeographic split of the Maritime Provinces breeding population with the mainland (Colbeck et al., 2008), conservation of this migratory population may also be important for preserving genetic diversity within the species. The Western Boreal population, ranging from Alaska to Saskatchewan, was characterized by the demographic analysis as having the third highest abundance and density, with population declines larger than the range-wide decline. Strong migratory connectivity with Mexico and Central America highlights the need for conservation efforts to focus on the most western portion of the range for this migratory population.

The Southern Temperate breeding population is unique in American Redstarts, in that nonbreeding individuals were sampled in both the far eastern Caribbean as well as in Central America. Our lack of sampling between these regions in northern South America precludes our ability to describe whether there is a migratory divide within the Southern Temperate population, or individuals are spread across this portion of the nonbreeding range. While weak connectivity across a large nonbreeding distribution could promote resilience from stressors on any single portion of the nonbreeding distribution (Finch et al., 2017), this makes targeting regions for conservation difficult.

*Low-coverage WGS for population assignment*

In addition to elucidating fine-scale migratory connectivity patterns in the American Redstart, our results provide important considerations for other population assignment studies using lcWGS. We found that balancing *effective sample sizes* of the source populations to within one effective individual of each other was essential for accurate assignment. Even when the actual number of individuals used per population was the same, variation in mean depth (1.3X – 1.9X) between populations skewed the effective sample sizes, resulting in decreased assignment accuracy. Other studies with known genotypes from RADseq have demonstrated the influence of actual sample size on overall assignment accuracy but not how it affects assignment bias (Benestan et al., 2015; DeSaix et al., 2019). The effective sample sizes needed per population for accurate assignment and the degree of standardizing these values will depend on the population structure of the study system. For example, study systems with higher genetic differentiation between populations may not need to finely standardize effective sample size to achieve high assignment accuracy. We suggest that other population assignment studies similarly evaluate the

influence of source population effective sample size on known source individuals before assigning individuals of unknown origin. Reducing the effective sample size of a sampled population can be achieved by either removing individuals or down sampling the read depth. In this study, we chose to remove individuals, and used the individuals' effective sample sizes as a guide for how many individuals to remove from each population (resulting in 21 – 27 samples per population). For studies with smaller sample sizes, it may be worthwhile to investigate if retaining all individuals but down sampling reads is a better alternative for standardizing effective sample sizes to retain more variation from individuals. Additionally, it may be important for studies with widespread admixture across populations to refine their sample selection for source populations by removing the more admixed individuals. Finally, studies with poor assignment accuracy to weakly differentiated populations may benefit from identifying a subset of informative loci based on elevated signatures of genetic differentiation (Ruegg et al., 2014; DeSaix et al., 2019). For example, our genome-wide $F_{ST}$ calculations reveal regions of elevated genetic differentiation that could be used to identify a smaller subset of SNPs for performing assignment. However, given our initial high assignment accuracy using the full set of SNPs ($n$ = 4,722,390) and the computational efficacy of WGSassign with these data, we suggest studies should first evaluate population assignment with genome-wide data to determine if the more labor-intensive process of informative SNP selection is warranted for their study system.

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

and the number of locations sampled) across nonbreeding stages of the annual cycle can drastically improve our understanding of population-level connectivity at low cost. Combined with cost-effective approaches for library preparation (e.g. Schweizer & DeSaix, 2023; Therkildsen & Palumbi, 2017), lcWGS is increasingly becoming economically feasible for a wide-range of studies. However, a trade-off with lcWGS is that the sequence data processing requires additional costs associated with time spent on the bioinformatics analysis. For studies interested in population assignment with a large number of samples, increasing the number of samples per lane, thereby decreasing the mean average sequencing depth, may make lcWGS economically feasible compared to other sequencing methods. For a comprehensive review of coverage guidelines for different types of analyses with low-coverage WGS data see Lou et al. (2021).

An interesting aspect of our results was that all posterior probabilities of assignment were > 0.8, even for potentially admixed individuals. A standard method to determine assignment confidence in population assignment studies is to use a cutoff value for posterior probabilities of assignment (DeSaix et al., 2019; Ruegg et al., 2014). Individuals with low posterior probabilities of assignment (e.g., < 0.8) can be highly admixed. Thus, it is inaccurate to classify them as from a specific population. However, we suspect that with lcWGS data, the high prevalence of loci with single read results in the likelihood being highest for a homozygous genotype. Thus, admixed individuals may "switch" their population of maximum likelihood depending on the loci used for assignment. Our use of an assignment consistency threshold addressed this concern by creating subsets of genomic data for population assignment to determine if individuals could reliably be assigned to a single population when different loci were used. Testing the assignment consistency threshold with known source individuals revealed three individuals with inconsistent

assignment (< 0.8, i.e., 8 out of 10 genomic datasets) and were likely admixed between pure Northern Temperate and Southern Temperate populations. These results highlight that the consistency of assignment may be more reliable than posterior probabilities for confidently assigning individuals of unknown origin. Further development of spatially explicit assignment methods for genotype likelihood data would be helpful for determining the likely origin of admixed individuals at the periphery of source populations.

*Conclusion*

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

## Tables and figures

**Table 3.1.** Assignment accuracy of the known breeding samples used as the testing set (n = 47) for the effective size standardized breeding populations. Down sampling to both 0.1X and 0.01X achieved high assignment accuracy.

| Depth | Western Boreal | Basin Rockies | Southern Temperate | Northern Temperate | Total |
|---|---|---|---|---|---|
| Full | 100% (5/5) | 100% (6/6) | 100% (22/22) | 93% (13/14) | 98% (46/47) |
| 0.1X | 100% (5/5) | 100% (6/6) | 100% (22/22) | 100% (14/14) | 100% (47/47) |
| 0.01X | 100% (5/5) | 100% (6/6) | 95% (21/22) | 100% (14/14) | 98% (46/47) |

**Table 3.2.** Demographic analysis of Breeding Bird Survey data for breeding populations of American Redstart.

| Population | No. of BBS routes | Average trend (95% CI) | Relative abundance index (95% CI) | Route density (95% CI) |
|---|---|---|---|---|
| Western Boreal | 228 | -0.90 (-1.54, -0.27) | 0.66 (0.52, 0.84) | 1.67 (1.31, 2.11) |
| Basin Rockies | 84 | -2.38 (-3.40, -1.30) | 0.01 (0.01, 0.02) | 0.16 (0.10, 0.25) |
| Southern Temperate | 673 | -0.34 (-0.75, 0.06) | 0.15 (0.13, 0.17) | 0.67 (0.58, 0.78) |
| Northern Temperate | 615 | 0.67 (0.33, 1.01) | 3.70 (3.10, 4.52) | 15.16 (12.68, 18.54) |
| Maritime Provinces | 166 | -1.26 (-1.73, -0.77) | 1.97 (1.57, 2.49) | 30.56 (24.33, 38.60) |

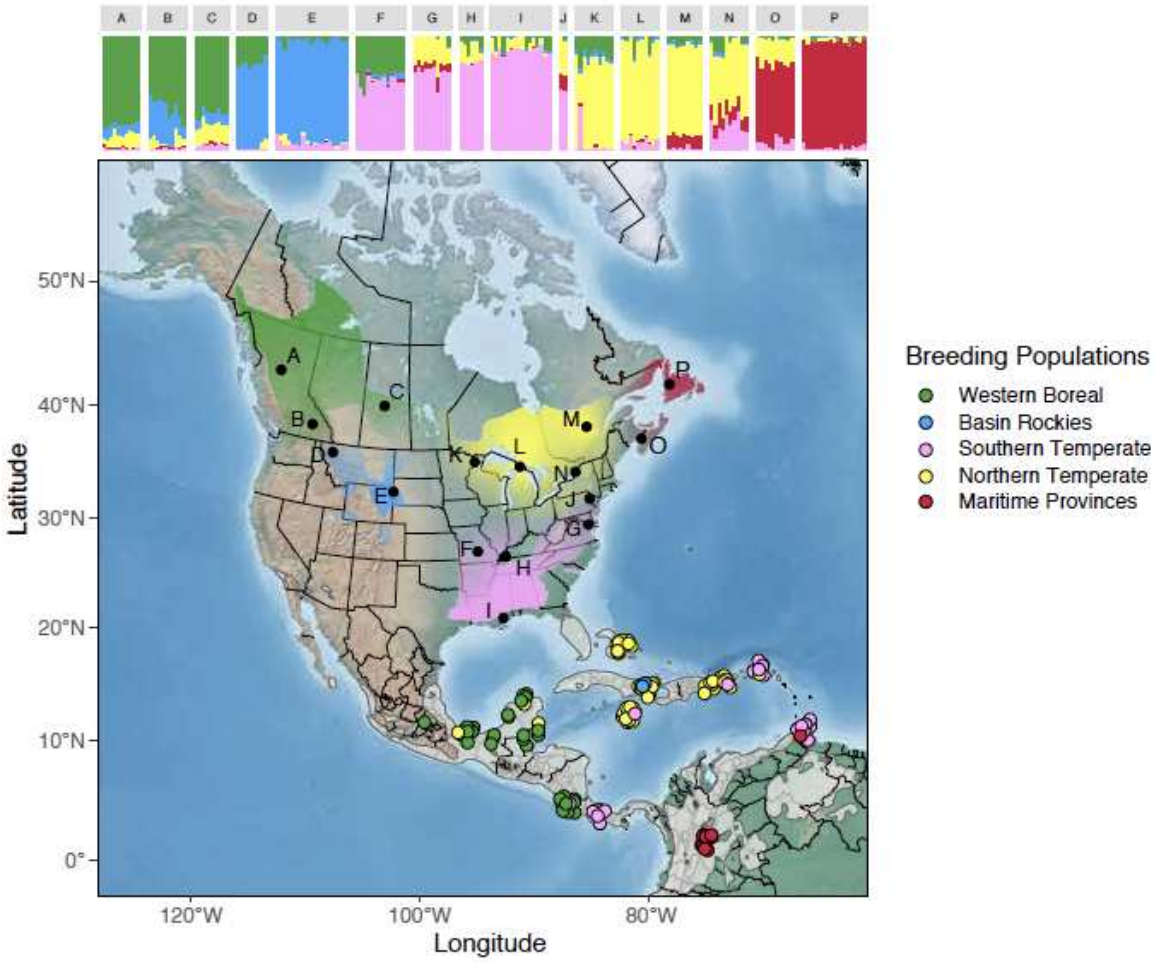


**Figure 3.1.** The population structure on the breeding range is delineated by five genetically distinct clusters (colored polygons) from the results of the admixture analysis (top panel). Population structure was determined using 169 individuals from 16 sites (black points) on the breeding range. Individuals sampled from the non-breeding range were determined to originate from a given breeding population through population assignment tests (n = 138; colored circles). The point colors on the nonbreeding range represent the breeding population of maximum likelihood of assignment and the extent of the nonbreeding range is provided by the grey polygon. Strong migratory connectivity is evident from the general separation of breeding population assignment across the wintering range.

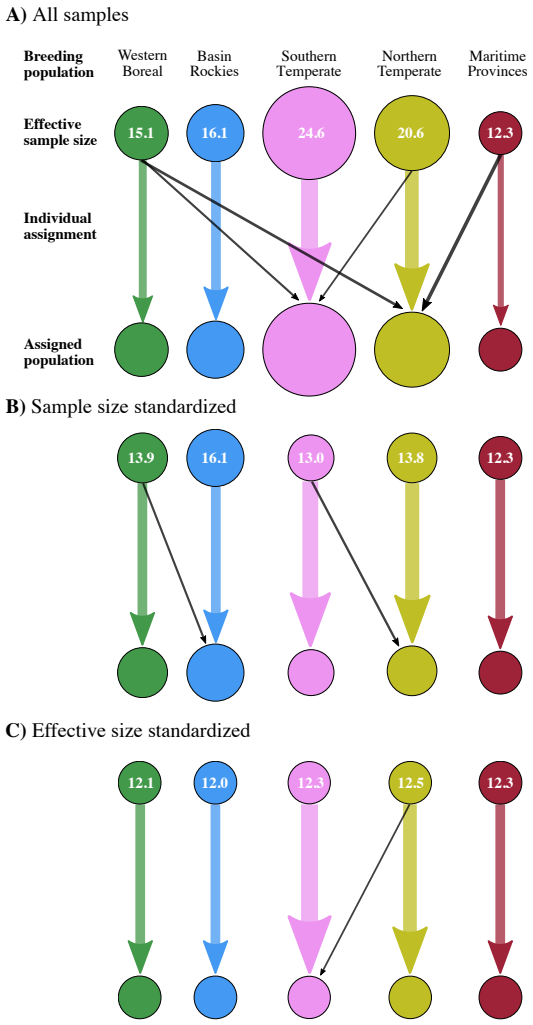


**Figure 3.2.** Population assignment of known breeding individuals revealed assignment bias from unequal effective sample sizes. Circles represent the breeding populations (colored), with circle size representing effective sample size, and arrows represent the assignment of individuals from their known breeding population to their assigned population. Arrows are scaled in size by the number of individuals assigned. Colored arrows represent the correct individuals assigned to a breeding population, whereas black arrows indicate incorrect assignment to a different population. A) When using all samples to calculate allele frequencies in breeding populations, all incorrectly assigned individuals (n = 18) were assigned to a population with higher effective sample size. B) Standardizing breeding populations by sample sizes (27 individuals per population) resulted in less incorrect assignment (n = 5), but all individuals were still assigned to another population with higher effective sample size. C) Standardizing breeding populations to approximately the same effective sample size (~12 effective individuals), resulted in only one individual being incorrectly assigned. In all cases, incorrect assignment was typically to a geographically neighboring population.

# LITERATURE CITED


Battey, C. J., Linck, E. B., Epperly, K. L., French, C., Slager, D. L., Sykes, P. W., & Klicka, J. (2018). A migratory divide in the painted bunting (Passerina ciris). American Naturalist, 191(2), 259–268. https://doi.org/10.1086/695439

Benestan, L., Gosselin, T., Perrier, C., Sainte-Marie, B., Rochette, R., & Bernatchez, L. (2015). RAD genotyping reveals fine-scale genetic structuring and provides powerful population assignment in a widely distributed marine species, the American lobster (Homarus americanus). Molecular Ecology, 24(13), 3299–3315. https://doi.org/10.1111/MEC.13245

Buerkle, A. C., & Gompert, Z. (2013). Population genomics based on low coverage sequencing: how low should we go? Molecular Ecology, 22(11), 3028–3035. https://doi.org/10.1111/MEC.12105

Colbeck, G. J., Gibbs, H. L., Marra, P. P., Hobson, K., & Webster, M. S. (2008). Phylogeography of a widespread North American migratory songbird (Setophaga ruticilla). Journal of Heredity, 99(5), 453–463. https://doi.org/10.1093/jhered/esn025

Corrales, L., Bouroncle, C., & Zamora, J. C. (2015). An overview of forest biomes and ecoregions of Central America. Climate change impacts on tropical forests in Central America, 33-54.

DeSaix, M. G., Rodriguez, M. D., Ruegg, K. C., & Anderson, E. C. (in review). Population assignment from genotype likelihoods for low-coverage whole-genome sequencing data. Molecular Ecology Resources. https://doi.org/10.22541/au.168569102.27840692/v1

DeSaix, M. G., Bulluck, L. P., Eckert, A. J., Viverette, C. B., Boves, T. J., Reese, J. A., Tonra, C. M., & Dyer, R. J. (2019). Population assignment reveals low migratory connectivity in a weakly structured songbird. Molecular Ecology, 28(9), 2122–2135. https://doi.org/10.1111/MEC.15083

DeSaix, M. G., Connell, E. B., Cortes-Rodríguez, N., Omland, K. E., Marra, P. P., & Studds, C. E. (2022). Migratory connectivity in a Newfoundland population of the American Redstart (Setophaga ruticilla). The Wilson Journal of Ornithology, 134(3), 381–389. https://doi.org/10.1676/22-00004

[dataset] DeSaix, M. G., Anderson, E. C., Bossu, C. M., Rayne, C. E., Schweizer, T. M., Bayly, N. J., Narang, D. S., Hagelin, J. C., Gibbs, H. L., Sarraco, J. F., Sherry, T. W., Webster, M. S., Smith, T. B., Marra, P. P., Ruegg, K. C; 2023; Low-coverage whole genome sequencing for highly accurate population assignment: Mapping migratory connectivity in the American Redstart (Setophaga ruticilla); Dryad; DOI PROVIDED BY MOLECULAR ECOLOGY

Faaborg, J., Holmes, R. T., Anders, A. D., Bildstein, K. L., Dugger, K. M., Gauthreaux, S. A., Heglund, P., Hobson, K. A., Jahn, A. E., Johnson, D. H., Latta, S. C., Levey, D. J., Marra, P. P., Merkord, C. L., Erica, N. O. L., Rothstein, S. I., Sherry, T. W., Scott Sillett, T., Thompson, F. R., & Warnock, N. (2010). Recent advances in understanding migration systems of New World land birds. Ecological Monographs, 80(1), 3–48. https://doi.org/10.1890/09-0395.1

Finch, T., Butler, S. J., Franco, A. M. A., & Cresswell, W. (2017). Low migratory connectivity is common in long-distance migrant birds. Journal of Animal Ecology, 86(3), 662–673. https://doi.org/10.1111/1365-2656.12635

Fraser, K. C., Stutchbury, B. J. M., Kramer, P., Silverio, C., Barrow, J., Newstead, D., Mickle, N., Shaheen, T., Mammenga, P., Applegate, K., Bridge, E., & Tautin, J. (2013). Consistent Range-Wide Pattern in Fall Migration Strategy of Purple Martin (Progne subis), Despite Different Migration Routes at the Gulf of Mexico. The Auk, 130(2), 291–296. https://doi.org/10.1525/AUK.2013.12225

Fumagalli, M., Vieira, F. G., Korneliussen, T. S., Linderoth, T., Huerta-Sánchez, E., Albrechtsen, A., & Nielsen, R. (2013). Quantifying population genetic differentiation from next-generation sequencing data. Genetics, 195(3), 979–992. https://doi.org/10.1534/GENETICS.113.154740/-/DC1

Garcia-Perez, B., & Hobson, K. A. (2014). A multi-isotope (δ2H, δ13C, δ15N) approach to establishing migratory connectivity of Barn Swallow (Hirundo rustica). Ecosphere, 5(2), 1–12. https://doi.org/10.1890/ES13-00116.1

Gelman, Andrew., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. In Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Gómez, C., Hobson, K. A., Bayly, N. J., Rosenberg, K. V., Morales-Rozo, A., Cardozo, P., & Cadena, C. D. (2021). Migratory connectivity then and now: a northward shift in breeding origins of a long-distance migratory bird wintering in the tropics. Proceedings of the Royal Society B, 288(1948). https://doi.org/10.1098/RSPB.2021.0188

González-Prieto, A. M., Bayly, N. J., Colorado, G. J., & Hobson, K. A. (2017). Topography of the Andes Mountains shapes the wintering distribution of a migratory bird. Diversity and Distributions, 23(2), 118–129. https://doi.org/10.1111/DDI.12515

Grant, W. S., Milner, G. B., Krasnowski, P., & Utter, F. M. (1980). Use of Biochemical Genetic Variants for Identification of Sockeye Salmon (Oncorhynchus nerka) Stocks in Cook Inlet, Alaska. Canadian Journal of Fisheries and Aquatic Sciences, 37(8), 1236–1247. https://doi.org/10.1139/F80-159

Hanghøj, K., Moltke, I., Andersen, P. A., Manica, A., & Korneliussen, T. S. (2019). Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. GigaScience, 8(5), 1–9. https://doi.org/10.1093/GIGASCIENCE/GIZ034

Kellner, K., & Meredith, M. (2021). jagsUI: A Wrapper Around “rjags” to Streamline “JAGS” Analyses (1.5.2). https://CRAN.R-project.org/package=jagsUI

Kimura, M., Clegg, S. M., Lovette, I. J., Holder, K. R., Girman, D. J., Milá, B., Wade, P., & Smith, T. B. (2002). Phylogeographical approaches to assessing demographic connectivity between breeding and overwintering regions in a Nearctic−Neotropical warbler (Wilsonia pusilla). Molecular Ecology, 11(9), 1605–1616. https://doi.org/10.1046/J.1365-294X.2002.01551.X

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics, 15(1), 1–13. https://doi.org/10.1186/S12859-014-0356-4/TABLES/4

Korneliussen, T. S., & Moltke, I. (2015). NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. Bioinformatics, 31(24), 4009–4011. https://doi.org/10.1093/BIOINFORMATICS/BTV509

Larison, B., Lindsay, A. R., Bossu, C., Sorenson, M. D., Kaplan, J. D., Evers, D. C., Paruk, J., DaCosta, J. M., Smith, T. B., & Ruegg, K. (2021). Leveraging genomics to understand threats to migratory birds. Evolutionary Applications, 14(6), 1646–1658. https://doi.org/10.1111/EVA.13231

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/BIOINFORMATICS/BTP324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, 1000 Genome Project Data Processing. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/BIOINFORMATICS/BTP352

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology, 30(23), 5966–5993. https://doi.org/10.1111/MEC.16077

Lou, R. N., & Therkildsen, N. O. (2022). Batch effects in population genomic studies with low-coverage whole genome sequencing data: Causes, detection and mitigation. Molecular Ecology Resources, 22(5), 1678–1692. https://doi.org/10.1111/1755-0998.13559

Lovette, I. J., Clegg, S. M., & Smith, T. B. (2004). Limited Utility of mtDNA Markers for Determining Connectivity among Breeding and Overwintering Locations in Three Neotropical Migrant Birds. Conservation Biology, 18(1), 156–166. https://doi.org/10.1111/J.1523-1739.2004.00239.X

Lovette, I. J., & Holmes, R. T. (1995). Foraging Behavior of American Redstarts in Breeding and Wintering Habitats: Implications for Relative Food Availability. The Condor, 97(3), 782–791. https://doi.org/10.2307/1369186

Marra, P. P., Cohen, E. B., Loss, S. R., Rutter, J. E., & Tonra, C. M. (2015). A call for full annual cycle research in animal ecology. Biology Letters, 11(8). https://doi.org/10.1098/RSBL.2015.0552

Marra, P. P., Hobson, K. A., & Holmes, R. T. (1998). Linking winter and summer events in a migratory bird by using stable- carbon isotopes. Science, 282(5395), 1884–1886. https://doi.org/10.1126/SCIENCE.282.5395.1884

Marra, P. P., & Holmes, R. T. (2001). Consequences of Dominance-Mediated Habitat Segregation in American Redstarts During the Nonbreeding Season. The Auk, 118(1), 92–104. https://doi.org/10.1093/AUK/118.1.92

Marra, P. P., Sherry, T. W., & Holmes, R. T. (1993). Territorial Exclusion by a Long-Distance Migrant Warbler in Jamaica: A Removal Experiment with American Redstarts (Setophaga Ruticilla). The Auk, 110(3), 565–572. https://doi.org/10.2307/4088420

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.Journal, 17(1), 10–12. https://doi.org/10.14806/EJ.17.1.200

Martin, T. G., Chadès, I., Arcese, P., Marra, P. P., Possingham, H. P., & Norris, D. R. (2007). Optimal Conservation of Migratory Species. PLOS ONE, 2(8), e751. https://doi.org/10.1371/JOURNAL.PONE.0000751

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303. https://doi.org/10.1101/GR.107524.110

Meisner, J., & Albrechtsen, A. (2018). Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. Genetics, 210(2), 719–731. https://doi.org/10.1534/GENETICS.118.301336

Newton, I. (2008). The migration ecology of birds. Academic Press.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y., & Wang, J. (2012). SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data. PLOS ONE, 7(7), e37558. https://doi.org/10.1371/JOURNAL.PONE.0037558

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 2011 12:6, 12(6), 443–451. https://doi.org/10.1038/nrg2986

Norris, D. R., Marra, P. P., Bowen, G. J., Ratcliffe, L. M., Royle, J. A., & Kyser, T. K. (2006). Migratory connectivity of a widely distributed songbird, the american redstart (Setophaga Ruticilla). Ornithological Monographs, 61, 14–28. https://doi.org/10.1642/0078-6594

Paetkau, D., Calvert, W., Sterling, I., & Strobeck, C. (1995). Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology, 4(3), 347–354. https://doi.org/10.1111/J.1365-294X.1995.TB00227.X

Pardieck, K. L., David, Z. J., Lutmerding, M., Aponte, V., & Hudson, M.-A. R. (2020). North American Breeding Bird Survey Dataset 1966—2019, version 2019.0 [Data set]. U.S. Geological Survey. https://doi.org/10.5066/P9J6QUF6

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. Working Papers.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ralston, J., FitzGerald, A. M., Burg, T. M., Starkloff, N. C., Warkentin, I. G., & Kirchman, J. J. (2021). Comparative phylogeographic analysis suggests a shared history among eastern North American boreal forest birds. Ornithology, 138(3). https://doi.org/10.1093/ORNITHOLOGY/UKAB018

Rannala, B., & Mountain, J. L. (1997). Detecting immigration by using multilocus genotypes. Proceedings of the National Academy of Sciences, 94(17), 9197–9201. https://doi.org/10.1073/PNAS.94.17.9197

Raybuck, D. W., Boves, T. J., Stoleson, S. H., Larkin, J. L., Bayly, N. J., Bulluck, L. P., ... & Buehler, D. A. (2022). Cerulean Warblers exhibit parallel migration patterns and multiple migratory stopovers within the Central American Isthmus. Ornithological Applications, 124(4), duac031.

Roberto-Charron, A., Kennedy, J., Reitsma, L., Tremblay, J. A., Krikun, R., Hobson, K. A., Ibarzabal, J., & Fraser, K. C. (2020). Widely distributed breeding populations of Canada warbler (Cardellina canadensis) converge on migration through Central America. BMC Zoology, 5(1), 1–14. https://doi.org/10.1186/S40850-020-00056-4/FIGURES/4

Rueda-Hernández, R., Bossu, C. M., Smith, T. B., Contina, A., Canales del Castillo, R., Ruegg, K., & Hernández-Baños, B. E. (2023). Winter connectivity and leapfrog migration in a migratory passerine. Ecology and Evolution, 13(2), e9769. https://doi.org/10.1002/ECE3.9769

Ruegg, K. C., Anderson, E. C., Paxton, K. L., Apkenas, V., Lao, S., Siegel, R. B., Desante, D. F., Moore, F., & Smith, T. B. (2014). Mapping migration in a songbird using high-resolution genetic markers. Molecular Ecology, 23(23), 5726–5739. https://doi.org/10.1111/MEC.12977

Ruegg, K. C., Harrigan, R. J., Saracco, J. F., Smith, T. B., & Taylor, C. M. (2020). A genoscape-network model for conservation prioritization in a migratory bird. Conservation Biology, 34(6), 1482–1491. https://doi.org/10.1111/COBI.13536

Ruegg, K. C., Hijmans, R. J., & Moritz, C. (2006). Climate change and the origin of migratory pathways in the Swainson's thrush, Catharus ustulatus. Journal of Biogeography, 33(7), 1172–1182. https://doi.org/10.1111/J.1365-2699.2006.01517.X

Ruegg, K. C., & Smith, T. B. (2002). Not as the crow flies: a historical explanation for circuitous migration in Swainson's thrush (Catharus ustulatus). Proceedings of the Royal Society of London. Series B: Biological Sciences, 269(1498), 1375–1381. https://doi.org/10.1098/RSPB.2002.2032

Rushing, C. S., Andrew Royle, J., Ziolkowski, D. J., & Pardieck, K. L. (2020). Migratory behavior and winter geography drive differential range shifts of eastern birds in response to recent climate change. Proceedings of the National Academy of Sciences of the United States of America, 117(23), 12897–12903. https://doi.org/10.1073/PNAS.2000299117

Rushing, C. S., Ryder, T. B., Saracco, J. F., & Marra, P. P. (2014). Assessing migratory connectivity for a long-distance migratory bird using multiple intrinsic markers. Ecological Applications, 24(3), 445–456. https://doi.org/10.1890/13-1091.1

Sauer, J. R., & Link, W. A. (2011). Analysis of the North American Breeding Bird Survey Using Hierarchical Models. The Auk, 128(1), 87–98. https://doi.org/10.1525/auk.2010.09220

Schweizer, T. M., & DeSaix, M. G. (2023). Cost-effective library preparation for whole genome sequencing with feather DNA. Conservation Genetics Resources. https://doi.org/10.1007/S12686-023-01299-2

Sherry, T. W., Holmes, R., Pyle, P., & Patten, M. (2016). American Redstart (Setophaga ruticilla). In P. G. Rodewald (Ed.), The Birds of North America Online. Ithaca: Cornell Lab of Ornithology.

Sherry, T. W., & Holmes, R. T. (1996). Winter Habitat Quality, Population Limitation, and Conservation of Neotropical-Nearctic Migrant Birds. Ecology, 77(1), 36–48.

Sillett, T. S., Holmes, R. T., & Sherry, T. W. (2000). Impacts of a global climate cycle on population dynamics of a migratory songbird. Science, 288(5473), 2040–2043. https://doi.org/10.1126/SCIENCE.288.5473.2040

Small-Lorenz, S. L., Culp, L. A., Ryder, T. B., Will, T. C., & Marra, P. P. (2013). A blind spot in climate change vulnerability assessments. Nature Climate Change 2013 3:2, 3(2), 91–93. https://doi.org/10.1038/nclimate1810

Strimas-Mackey, M., Ligocki, S., Auer, T., & Fink, D. (2022). ebirdst: Tools for loading, plotting, mapping and analysis of eBird Status and Trends data products. R package version 1.2021.0. https://cornelllabofornithology.github.io/ebirdst/

Studds, C. E., & Marra, P. P. (2011). Rainfall-induced changes in food availability modify the spring departure programme of a migratory bird. Proceedings of the Royal Society B: Biological Sciences, 278(1723), 3437–3443. https://doi.org/10.1098/RSPB.2011.0332

Studds, C. E., Wunderle, J. M., & Marra, P. P. (2021). Strong differences in migratory connectivity patterns among species of Neotropical-Nearctic migratory birds revealed by combining stable isotopes and abundance in a Bayesian assignment analysis. Journal of Biogeography, 48(7), 1746–1757. https://doi.org/10.1111/JBI.14111

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Molecular Ecology Resources, 17(2), 194–208. https://doi.org/10.1111/1755-0998.12593

Webster, M. S., Marra, P. P., Haig, S. M., Bensch, S., & Holmes, R. T. (2002). Links between worlds: unraveling migratory connectivity. Trends in Ecology & Evolution, 17(2), 76–83. https://doi.org/10.1016/S0169-5347

Wilson, S., Saracco, J. F., Krikun, R., Flockhart, D. T., Godwin, C. M., & Foster, K. R. (2018). Drivers of demographic decline across the annual cycle of a threatened migratory bird. Scientific Reports, 8(1), 7316.