Genetic Toolkit for Assessment and Prediction of Population-Level Impacts of Bridge Construction on Birds

January 2020

A Research Report from the National Center for Sustainable Transportation

Rachael Bay, University of California, Davis
Lisa A. Tell, University of California, Davis
Nicole Adams, University of California, Davis
Ruta R. Bandivadekar, University of California, Davis

National Center for Sustainable Transportation
ITS UCDAVIS INSTITUTE OF TRANSPORTATION STUDIES

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. NCST-UCD-RR-20-02	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Genetic Toolkit for Assessment and Prediction of Population-Level Impacts of Bridge Construction on Birds		**5. Report Date** January 2020 **6. Performing Organization Code** N/A
7. Author(s) Rachael Bay, Ph.D. http://orcid.org/0000-0002-9516-5881 Lisa A. Tell, D.V.M. http://orcid.org/0000-0003-1823-7420 Nicole Adams, Ph.D. https://orcid.org/0000-0001-7181-2477 Ruta R. Bandivadekar, M.V.Sc https://orcid.org/0000-0001-6380-8881		**8. Performing Organization Report No.** UCD-ITS-RR-20-08
9. Performing Organization Name and Address University of California, Davis Institute of Transportation Studies 1605 Tilia Street, Suite #100 Davis, CA 95616		**10. Work Unit No.** N/A **11. Contract or Grant No.** Caltrans 65A0686 Task Order 002 USDOT Grant 69A3551747114
12. Sponsoring Agency Name and Address California Department of Transportation Division of Research, Innovation and System Information, MS-83 1727 30th Street, Sacramento, CA 95816		**13. Type of Report and Period Covered** Final Report (December 2018–November 2019) **14. Sponsoring Agency Code** USDOT OST-R

15. Supplementary Notes
DOI: https://doi.org/10.7922/G2TX3CMT
Dataset DOI: https://doi.org/10.25338/B8W03P

16. Abstract
Recent studies have highlighted alarming rates of declines in bird populations across the country. The State of California is home to over 650 resident and migrant avian species. Legislation for protecting these species has existed for over a century now, yet tools for identifying populations and understanding seasonal movement remain limited. Recently, genetic and genomic tools have provided a method for understanding population structure, allowing for more informed delineation of management units. The goal of this project was to create a genetic toolkit for identifying breeding populations and assigning individuals to those populations. Ultimately, such tools could be used to assess population-level impacts when there are conflicts with birds at infrastructure construction sites. As a test case, we sequenced entire genomes for 40 individual Anna's hummingbirds (*Calypte anna*) from across the state. Based on this initial data, we found low levels of differentiation between sampled locations, suggesting that *C. anna* in California are not subdivided into different population units. However, there was a weak signal of geography suggesting there may be localized genetic differences in a small proportion of the genome. Follow-up work will focus on a broader sampling across the state of California to clarify any possible population subdivision or geographical patterns of differentiation.

17. Key Words Birds, Population Structure, Connectivity	**18. Distribution Statement** No restrictions.		
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 29	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

NCST

About the National Center for Sustainable Transportation

The National Center for Sustainable Transportation is a consortium of leading universities committed to advancing an environmentally sustainable transportation system through cutting-edge research, direct policy engagement, and education of our future leaders. Consortium members include: University of California, Davis; University of California, Riverside; University of Southern California; California State University, Long Beach; Georgia Institute of Technology; and University of Vermont. More information can be found at: ncst.ucdavis.edu.

Disclaimer

Acknowledgments

This study was funded, partially or entirely, by a grant from the National Center for Sustainable Transportation (NCST), supported by USDOT through the University Transportation Centers program. The authors would like to thank the NCST and USDOT for their support of university-based research in transportation, and especially for the funding provided in support of this project. The authors also thank Michael Clark and Brenda Cameron for assistance with lab work.

Genetic Toolkit for Assessment and Prediction of Population-Level Impacts of Bridge Construction on Birds

A National Center for Sustainable Transportation Research Report

January 2020

Rachael Bay, Department of Evolution and Ecology, University of California, Davis

Lisa A. Tell, School of Veterinary Medicine: Department of Medicine and Epidemiology, University of California, Davis

Nicole Adams, Department of Evolution and Ecology, University of California, Davis

Ruta R. Bandivadekar, School of Veterinary Medicine: Department of Medicine and Epidemiology, University of California, Davis

NCST

[page intentionally left blank]

TABLE OF CONTENTS

List of Tables

List of Figures

Genetic Toolkit for Assessment and Prediction of Population-Level Impacts of Bridge Construction on Birds

EXECUTIVE SUMMARY

Recent studies have highlighted alarming rates of declines in bird populations across the country. The State of California is home to over 650 resident and migrant avian species. Legislation for protecting these species has existed for over a century now, yet tools for identifying populations and understanding seasonal movement remain limited. Recently, genetic and genomic tools have provided a method for understanding population structure, allowing for more informed delineation of management units. The goal of this project was to create a genetic toolkit for identifying breeding populations and assigning individuals to those populations. Ultimately, such tools could be used to assess population-level impacts when there are conflicts with birds at infrastructure construction sites. As a test case, we sequenced entire genomes for 40 individual Anna's hummingbirds (*Calypte anna*) from across the state. Based on this initial data, we found low levels of differentiation between sampled locations, suggesting that *C. anna* in California are not subdivided into different population units. However, there was a weak signal of geography suggesting there may be localized genetic differences in a small proportion of the genome. Follow-up work will focus on a broader sampling across the state of California to clarify any possible population subdivision or geographical patterns of differentiation.

Introduction

California is home to nearly 650 species of birds, and its coastline and interior serve as vital breeding habitats and migration corridors for avian populations across the United States and Canada (DeSante 1983). Over the last century local, national, and international efforts to limit declines in bird populations have resulted in a number of legal protections, including the Migratory Bird Treaty Act and the California Fish and Game Code, which now protect most bird species in the state of California ("Laws Protecting Birds"; Golden Gate Audubon Society). Efforts to minimize impacts on birds during infrastructure projects such as bridge construction can cost millions of dollars and result in long-term delays (Cabanatun 2017; Associated Press 2017). In order to optimize planning of infrastructure we must understand the potential population-specific effects on wildlife populations.

A central challenge in the effort to achieve these goals is the lack of precise knowledge of population structure, including when and where particular populations are likely to be occupying a given area. More specifically, when a bird is found at a project site, knowledge of whether the individual is from a stable or declining population could have very different implications for resulting mitigation strategies. High-resolution maps of population structure are therefore the first step for understanding potential population-level impacts of construction projects on bird populations.

The effort to understand population-specific bird nesting and migration patterns has been met with numerous limitations and technological hurdles over the last century. Past efforts have focused on the recovery of individually banded birds, but recapture rates are extremely low (Bridge et al., 2013; Arlt et al., 2013; Kelly et al., 2005). More recently, geo-locators have been used to track songbird movement, but remain impractical for large-scale (1000s of individuals) applications due to cost, weight restrictions, and the need to recover individuals to collect data from the devices (Bridge et al., 2013; Kelly et al., 2005). Alternatively, genetic and isotopic markers that use information contained within a single feather to pinpoint an individual's population of origin have broad appeal because they are cost-effective, noninvasive, and do not require recapture (Kelly et al., 2005; Rubenstein et al., 2002). However, until recently these methods yielded poor resolution data and were plagued by technical issues related to working with feather material (Segelbacher 2002).

We avoid many issues that plagued previous methods using a recently-developed high-resolution genetic tagging method (Ruegg et al., 2014). We first use information across the entire genome of a bird to map genetic variation within a species across geographic space at finer spatial scales than previously possible. Once this map is established, we can then use genomic information contained within a single feather of a living bird or carcass to identify the breeding population of origin. Because of the low per-individual cost of screening, the high-resolution molecular tags that we have developed can be used to screen thousands of samples. This general framework can be used to delineate different populations across the breeding range as well as connect breeding, migratory, and wintering regions. For example, this method was recently used to define previously unresolved populations of the neotropical migratory

songbird Wilson's warbler and to define the Central and South American wintering range for each population (Ruegg et al., 2014).

The goal of our project was to develop a genetic toolkit that can be used to identify breeding populations for Anna's hummingbird. California comprises much of the native range of Anna's and it is abundant across much of the state. Although the species overall does not appear to be in decline, demographics vary across the range (Battey 2019). In recent decades, Anna's have shown a drastic range expansion from their previous range edge in northern California and birds can now be found wintering in British Columbia, likely assisted by resources provided by ornamental plants and birdfeeders. Currently, we know little about population structure, including the number, spatial extent, and status (stable or declining) of each population. Anna's Hummingbirds found nesting at bridge construction sites have led to multi-month delays (Associated Press 2017) despite the fact that we lack information about the potential impacts of such conflicts on population viability. In this report, we: 1) analyze population structure in Anna's hummingbirds across their California range and 2) Identify genetic variation most closely associated with geography.

Methods

Sample Collection

Hummingbirds were trapped using previously published methods (Russell and Russell, 2001) by permitted hummingbird banders (#23947). Blood was collected (20-30 μl, less than 1 % body weight), and placed in Queen's lysis buffer (Seutin et. al., 1991). All collection methods were approved by the University of California, Davis IACUC (F#20355) and California Department of Fish and Wildlife (permit SC-13066). For population structure analysis it is vital that birds are caught in their breeding location, therefore females were checked for visual signs of breeding including distension of cloaca with/without wrinkled skin, visualization of enlarged oviduct or egg in the coelomic cavity, and/or nesting behavior like presence of spider webbing on the beak of the female, sheared tail feather tips etc. In total, we collected 177 birds for this study. We added to this collection an additional 91 birds from museum collections and wildlife centers (Figure 1).

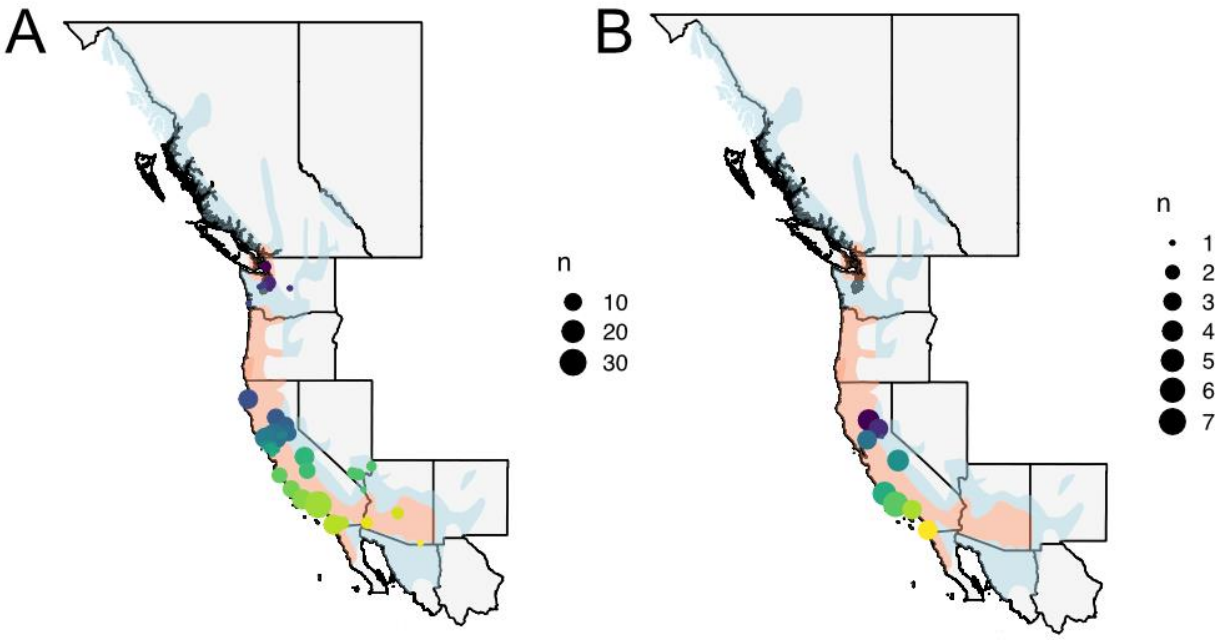

Figure 1. Anna's hummingbird samples collected for genomic analysis. Left panel (A) shows all samples including newly collected and retrieved from museums and wildlife centers. Right panel (B) shows locations of individuals used for whole genome sequencing.

DNA Extraction

Whole genomic DNA was extracted from 100-150µl of blood and lysis buffer mixture using the DNeasy Blood & Tissue Kit (Qiagen). The following modifications to the extraction protocol were used: samples were incubated overnight at 56°C, the sample was passed over the spin column twice prior to washing, an extra column drying step was taken (14000rpm for 3min), and DNA was eluted in 200µl AE buffer heated to 56°C. Whole genomic DNA was quantified using a Qubit Fluorometer (Thermo Fisher Scientific) and the quality of DNA was assessed using a 2% agarose gel.

Library Preparation and Sequencing

We used a modified library preparation based on Illumina's Nextera protocol (Baym et al., 2015; Overgaard Therkildsen and Palumbi, 2017) to sequence entire genomes of 40 birds (Figure 1B; Appendix A). To start, genomic DNA was standardized to 3ng/µl then underwent a tagmentation step using TDE1 enzyme and buffer (Illumina). Dual combination Nextera indexes (Illumina) were then added to tagged DNA fragments followed by a reconditioning step using the Kapa HiFi Kit (Kapa Biosystems). Libraries were then double size selected using AMPure XP Beads (Beckman Coulter) and quantified using a Qubit Fluorometer (Thermo Fisher Scientific). Sixteen libraries only went through a left side selection.

All libraries were pooled equimolarly then visualized with a Bioanalyzer (Agilent). The pooled libraries were further size selected to 320-600bp fragments using a Blue Pippin (Sage Science) at the DNA Technologies and Expression Analysis Core at the UC Davis Genome Center. The final library was sequenced on an Illumina HiSeq4000 with PE150 and the resulting sequences were demultiplexed by Novogene (Sacramento, CA, USA). All raw data can be accessed on DRYAD.

Sequence Processing and Analyses

Duplicate reads were removed using FastUnique (Xu et al., 2012), adapters and low-quality reads were trimmed using Trimmomatic (Bolger et al., 2014), and overlapping read pairs were collapsed into single reads using Flash (Magoc and Salzberg, 2011). Each sample was aligned to the *Calypte anna* reference genome (GCA_003957555.2, Korlach et al., 2017) using bwa (Li and Durbin, 2009) then sorted and indexed using Samtools (Li et al., 2009). Genetic variants or single nucleotide polymorphisms (hereafter called SNPs) were identified and genotype likelihoods were estimated using the ANGSD tool (Korneliussen et al., 2014) accessed through NGStools (Fumagalli et al., 2014) using the following parameters: -trim 0 -maxDepth 500, -minMapQ 20 -minQ 30 -minInd 20 -doCounts 1, -GL 1 -doMajorMinor 1 -doMaf 1 -skipTriallelic 1, -SNP_pval 1e-6, -minMaf 0.03. And whole genome coverage was calculated using Samtools.

Population structure was analyzed using principle components as well as hierarchical clustering analysis. A covariance matrix was calculated using PCAngsd (Meisner and Albrechtsen, 2018) and used in RStudio (RStudio Team, 2018) with R (R Core Team, 2019) to conduct eigenvector decomposition and create biplot comparing the first two principal components axes (PCs). A standard linear model was run to determine which factors contributed to PC axes. We ran separate linear models for each PC axis (as the response variable) and included sampling location, sex, and library prep date as explanatory variables. We used clustering in NGSadmix to infer the “best” number of populations and estimate ancestry proportions. We ran NGSadmix 10 times each with population numbers (K values) ranging from 1 to 5. We used the Evanno method implemented in CLUMPAK (Kopelman et al., 2015) to determine the best fit K value.

Design and Testing of Population Assignment Essay

To test the ability of a targeted SNP assay to assign a bird to its breeding location, we identified variants that showed greatest differentiation between the 5 northern and 4 southern sites. We used realSFS, part of NGSTools (Fumagalli et al., 2004) to estimate F_{ST} (population differentiation) for each site with minor allele frequency >0.1. We then ranked the sites by F_{ST} and thinned that list, removing sites within 1000 bp of one another to minimize redundancy due to linkage disequilibrium. We then took the top 100, 200, and 500 SNPs to create *in silico* assays. We use 100 SNPs as a minimum because this is the size that would easily be genotyped using as Fluidigm SNPtype assay, as in Ruegg et al. (2014). We recalculated PCAs using these SNP subsets to determine whether these assays would have the power to differentiate populations in different regions of California.

Results

Sequencing Output and SNP Identification

Across 40 individuals, our sequencing run produced 480 million short reads. Reads were high quality, with >95% of reads having quality scores Q>30 (99.9% base call accuracy) for all individuals. We obtained high variance in coverage among samples, with individual coverage ranging from 0.33X to 5.68X. Based on alignment to the *Calypte anna* reference genome, we were able to identify 11,333,093 SNPs.

Population Structure

Based on a preliminary PCA plot, we removed four obvious outliers from downstream analysis (Appendix B). When the PCA was recalculated for the remaining individuals, we observed weak separation between the 4 southern and 5 northern (see Figure 1B) locations (Figure 2) along the first principle component (3.12% of variance). Some of this signal may be confounded by a signal for library preparation; a standard linear model shows significant effects of library preparation date ($p=0.0002$) as well as sampling location ($p<0.0001$). PC2 (not shown) separated three individuals sampled in San Diego from all other samples. PC3 separated males from females, though this axis explained a lower fraction of the variance (2.90%). PC3 also had significant effects of both sampling location ($p<0.0001$) and sex ($p<0.0001$), but not of library preparation date ($p=0.71$).

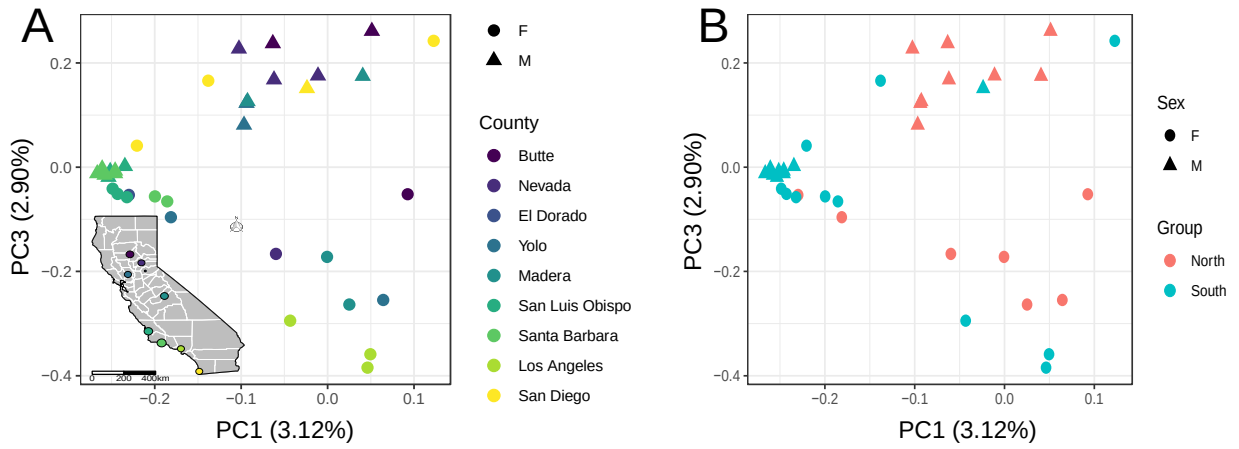

Figure 2. Principle components analysis (PCA) of genetic variation across the Anna's hummingbird genome. (A) Colored by location; colors match those from Figure 1 and (B) Colored based on north/south groupings, with the northern 5 locations in pink and the southern 4 locations coded in blue.

Hierarchical clustering analysis using NGSadmix shows similarly weak population structure across the California range of *C. anna*. Based on the Evanno method, the "optimal" number of clusters is K=3, suggesting there could be three separate population clusters within California. However, a look at the ancestry proportions of these three clusters does not produce any obvious geographic signal (Figure 3). Methods for choosing the based K are known to be biased,

and caution should be used when interpreting these values (Janes et al. 2017). Based on this data, we therefore conclude that there are no obvious barriers to gene flow between these locations for *C. anna*.

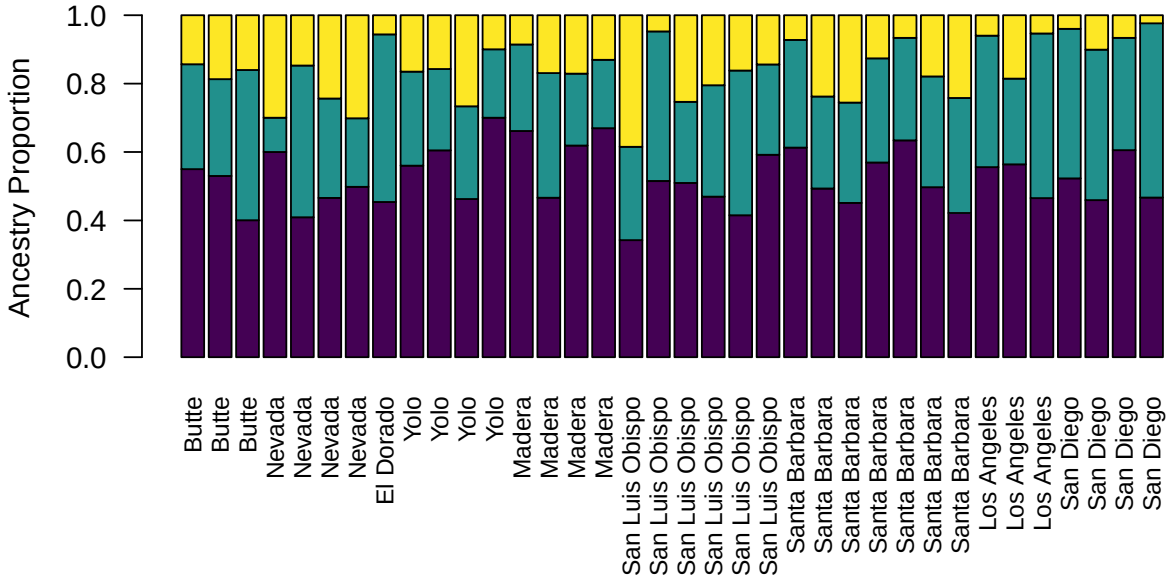

Figure 3. Ancestry proportions from hierarchical clustering analysis of 36 *C. anna* individuals. Here we visualize the mean ancestry proportions across 10 separate runs of NGSAdmix with K=3.

Testing of SNPs for Population Assignment

Overall, F_{ST} estimates between northern and southern locations were very low (global F_{ST}=0.003), supporting findings of little differentiation based on the population structure analysis. There were, however, a number of SNPs with higher levels of differentiation. We tested assays of three separate sizes: 100 SNPs (mean F_{ST}=0.32), 200 SNPs (mean F_{ST}=0.29), and 500 SNPs (mean F_{ST}=0.24). Principal component biplots for these three assays are shown in Figure 4. All assays fully differentiated northern from southern populations, suggesting few SNPs are necessary to assign birds of unknown origin to a breeding region. We did not gain any resolution by using 500 SNPs (Figure 4C) compared to 100 SNPs (Figure 4A). The 100 SNP assay is listed in Appendix C.

We discovered one region of the genome with an "F_{ST} peak", or region with many differentiated SNPs adjacent to one another (Appendix D). Although further testing is needed for validation, this region could be explored as an even more simple diagnostic tool to assign birds to geographical locations.

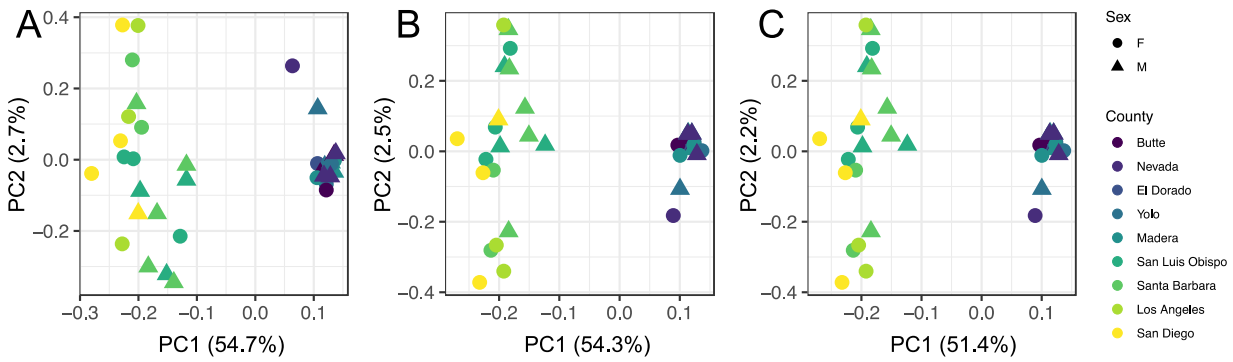

Figure 4. Principal components biplot of SNPs that differentiate northern from southern *C. anna*. Here we show assays of three different sizes: A) 100 SNPs, B) 200 SNPs, and C) 500 SNPs

Discussion

Understanding population structure is an important first step for managing wildlife populations (Funk et al., 2012; Palsbøll et al., 2007). Knowledge of population boundaries is vital to a wide array of conservation goals including estimating population declines (Segelbacher & Storch 2002; Moritz et al., 1997), identifying sources of adaptive variation (Funk et al., 2019), measuring gene flow and connectivity (Segelbacher et al., 2010), and estimating inbreeding (Darvill et al., 2006). Because infrastructure projects can directly impact genetic diversity and gene flow, development of a toolkit with which to assess population structure and the distribution of genetic variation could aid in predicting the impacts of such projects and wildlife. Here, we focus on the development of such a toolkit in the species *C. anna*, which has most of its range in California and has been the focus of previous conflicts with infrastructure projects.

Prior to this, little was known about genetic variation, population structure, and movement in *C. anna.* Prior to the early 1900s, the northern range limit was around the Sacramento Valley (Grinnell 1915), but over the last century that range has expanded and individuals can be found breeding as far north as British Columbia (Scarfe & Finlay 2001; Battey 2019). Population dynamics vary across this range, with winter population sizes in central California maintaining constant sizes over the last half century and populations in northern California increasing (Battey 2019). This difference in demography led us to wonder whether there were barriers to gene flow that were separating populations with different demographic trajectories or whether these differences were simply a product of a constantly expanding range.

Based on our analysis of 40 whole genomes gathered from across the state of California, we find little evidence of barriers to gene flow. We do find weak geographic differences among sampling locations; principle components analysis showed significant differences between birds sampled from different regions. However, our clustering analysis did not reveal any obvious geographical patterns, so any differentiation between sampled populations is likely weak. One issue with our data is that we see a potential artifact of library preparation. Such batch effects are common in next generation sequencing data and can obscure true biological signals (Taub

et al., 2010). More extensive sequencing across the range in addition to validation through alternative sequencing methods will be needed to eliminate the potential batch effects on inference of population structure.

Although population structure analysis revealed few barriers to gene flow, we were able to identify some genetic variants that distinguished between geographical locations. We tested the power of 100, 200, and 500 SNP assays to distinguish between northern and southern populations. All assays showed complete differentiation between these groups suggesting the ability to design assays to assign geographical origin despite apparent weak population structure. We did observe one region of the genome with a peak in differentiation. Regions like this can be the result of natural selection acting to differentiate populations that live in different environments or as the result of reproductive isolating forces. Further validation across more populations should be done to determine whether this region alone could serve as a simple assay to distinguish northern from southern birds.

With the exception of invasive species, birds in California are protected by a number of legal frameworks, including the federal Migratory Bird Treaty Act as well as the state California Fish and Game Code. In the past, mitigations efforts costing millions of dollars have been undertaken and projects have been delayed for months, sometimes due to even a single bird nesting at a bridge construction site. Estimations of population structure can help identify the population of origin of such birds and estimate the potential negative effects of such conflict with infrastructure projects.

Follow Up

We find little population structure among the sequenced populations. However, a more complete view of whether multiple populations of *C. anna* exist in California would be facilitated by more extensive geographic sampling as well as resequencing of low coverage samples from this initial project. We are currently working to obtain data from more locations (see samples in Figure 1A) in order increase the resolution of population structure analysis. Results should lead to a more high-resolution definition of populations across the California range of *C. anna*.

References

Arlt, D., Low, M. & Pärt, T. Effect of geolocators on migration and subsequent breeding performance of a long-distance passerine migrant. 2013. *PLoS ONE*, 8, e82316.

Associated Press. “Tiny hummingbird egg stalls project to upgrade a Bay Area bridge” 2017. Los Angeles Times.

Battey, C. J. Ecological release of the Anna's Hummingbird during a northern range expansion. 2019. *The American Naturalist,* 194: 306:315.

Baym M, Kryazhimskiy S, Lieberman TD, Chung H, Desai MM, Kishony R Inexpensive multiplexed library preparation for megabase-sized genomes. 2014. *PLoS ONE*, 10: e0128036.

Bolger AM, Lohse M, Usadel B Trimmomatic: A flexible trimmer for Illumina sequence data. 2014. *Bioinformatics,* 30: 2114–2120.

Brewer, L. W., Redmond, C. A., Stafford, J. M. & Hatch, G. E. Marking ruby throated hummingbirds with radio frequency identification tags. 2011. *The Journal of Wildlife Management,* 75: 1664–1667.

Bridge, E. S., Kelly, J. F., Contina, A., Gabrielson, R. M., MacCurdy, R. B. & Winkler, D.W. Advances in tracking small migratory birds: a technical review of light-level geolocation. 2013. *Journal of Field Ornithology*, 84: 121–137.

Cabanatuan, M. “Bay Bridge bird colony settles in on new span.” 2017. San Francisco Chronicle. http://www.sfchronicle.com.

Darvill, B., Ellis, J. S., Lye, G. C., & Goulson, D. Population structure and inbreeding in a rare and declining bumblebee, *Bombus muscorum* (Hymenoptera: Apidae). 2006. *Molecular Ecology*, *15*: 601–611.

DeSante, D. F. Annual variability in the abundance of migrant landbirds on Southeast Farallon Island, California. 1983. *The Auk*, 826–852.

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

Funk, W. C., Forester, B. R., Converse, S. J., Darst, C., & Morey, S. Improving conservation policy with genomics: a guide to integrating adaptive potential into US Endangered Species Act decisions for conservation practitioners and geneticists. 2019. *Conservation Genetics*, *20*: 115–134.

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. Harnessing genomics for delineating conservation units. 2012. *Trends in Ecology & Evolution*, *27*: 489–496.

Golden Gate Audubon Society. “Laws Protecting Birds.” http://goldengateaudubon.org.

Grinnell, J. *A distributional list of the birds of California*. 1915. Cooper Ornithological Club.

Janes, J.K., Miller, J.M., Dupuis, J.R., Malenfant, R.M., Gorrell, J.C., Cullingham, C.I., Andrew, R.L. The K=2 conundrum. 2017. *Molecular Ecology, 26:* 3594–3602.

Kelly, J. F., Ruegg, K. C. & Smith, T. B. Combining isotopic and genetic markers to identify breeding origins of migrant birds. 2005. *Ecological Applications*, 15: 1487–1494.

Kopelman, N. M., Mayzel, J., Jakobsson, M., Rosenberg, N. A., & Mayrose, I. Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. 2015. *Molecular Ecology Resources*, *15*: 1179–1191.

Korlach J, Gedman G, Kingan SB, Chen-Shan Chin, Howard JT, Audet J-N, Cantin L, Jarvis ED. *De novo* PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. 2017. *GigaScience*, 6: gix085.

Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. 2014. *BMC Bioinformatics*, 15: 443.

Li H and Durbin R .Fast and accurate short read alignment with Burrows-Wheeler Transform. 2009. *Bioinformatics*, 25:1754–1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. The Sequence alignment/map (SAM) format and SAMtools. 2009. *Bioinformatics*, 25:2078–2079.

Magoc T and Salzberg S. FLASH: Fast length adjustment of short reads to improve genome assemblies. 2011. *Bioinformatics*, 27: 2957–2963.

Meisner J and Albrechtsen A. Inferring population structure and admixture proportions in low-depth NGS data. 2018. *Genetics*, 210:719–731.

Moritz, C., Heideman, A., Geffen, E., & McRae, P. Genetic population structure of the greater bilby Macrotis lagotis, a marsupial in decline. 1997. *Molecular Ecology*, *6*: 925–936.

Palsbøll, P. J., Berube, M., & Allendorf, F. W. Identification of management units using population genetic data. 2007. *Trends in Ecology & Evolution*, *22*: 11–16.

RStudio Team RStudio: Integrated Development for R. RStudio, Inc., Boston, MA. 2018. http://www.rstudio.com/.

R Core Team R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2019. https://www.R-project.org/.

Rubenstein, D. R., Chamberlain, C. P., Holmes, R. T., Ayres, M. P., Waldbauer, J. R., Graves, G. R. & Tuross, N. C. Linking breeding and wintering ranges of a migratory songbird using stable isotopes. 2002. *Science,* 295: 1062–1065.

Ruegg, K. C., Anderson, E. C., Paxton, K. L., Apkenas, V., Lao, S., Siegel, R. B., DeSante, D. F., Moore, F. & Smith, T. B. Mapping migration in a songbird using high-resolution genetic markers. 2014. *Molecular Ecology,* 23: 5726–5739.

Russell, S. M., & Russell, R. O. The North American banders' manual for banding hummingbirds. 2011. The North American Banding Council. Point Reyes Station, California, USA.

Scarfe, A., and J. C. Finlay. Rapid second nesting by Anna's hummingbird near its northern breeding limits. 2001. *Western Birds,* 32:131–133.

Segelbacher, G. Noninvasive genetic analysis in birds: testing reliability of feather samples. 2002. *Molecular Ecology Notes*, 2: 367–369.

Segelbacher, G., Cushman, S.A., Epperson, B.K., Fortin, M.J., Francois, O., Hardy, O.J., Holderegger, R., Taberlet, P., Waits, L.P. and Manel, S. Applications of landscape genetics in conservation biology: concepts and challenges. 2010. *Conservation genetics*, *11*: 375–385.

Segelbacher, G., & Storch, I. Capercaillie in the Alps: genetic evidence of metapopulation structure and population decline. 2002. *Molecular Ecology*, *11*: 1669–1677.

Seutin, G., White, B. N., Boag, P.T. Preservation of avian blood and tissue samples for DNA analyses. 1991. *Canadian Journal of Zoology,* 69:82:90.

Taub, M. A., Bravo, H. C., & Irizarry, R. A. Overcoming bias and systematic errors in next generation sequencing data. 2010. *Genome Medicine*, *2*: 87.

Therkildsen, N. O. & Palumbi, S. R. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. 2016. *Molecular Ecology Resources*, 17: 194–208.

Xu H, Luo X, Qian J, et al. FastUniq: a fast de novo duplicates removal tool for paired short reads. 2012. *PLoS ONE*, 7: e52249.

Data Management

Products of Research

We conducted whole genome sequencing for 40 birds sampled from across the state. Blood or tissue samples as well as DNA extracts are archived in the Tell Lab at UC Davis. Sequencing was performed on an Illumina HiSeq 4000 at Novogene (Sacramento, CA). We generated between 1.4 million and 25 million short reads per individual bird.

Data Format and Content

The raw data is archived in .fastq format, which provides base pair reads as well as sequencing quality scores. Metadata, including information on capture location and date is stored as a spreadsheet.

Data Access and Sharing

The data is publicly available on the Dryad data repository at https://doi.org/10.25338/B8W03P.

Reuse and Redistribution

Data is available for public use, providing it is properly referenced. Suggested reference:

Adams, Nicole et al. (2019), Low coverage whole genomes of Calypte anna across California, USA, v5, UC Davis, Dataset, https://doi.org/10.25338/B8W03P

Appendix A

Table A1. Metadata for individual *C. anna* used for whole genome sequencing. "Raw reads" refers to the total number of short reads from the sequencing run assigned to that sample. "Coverage" was calculated using samtools.

ID	Town	County	Sex	Raw reads	Coverage
ANHU_16592	Placerville	El Dorado	F	1850204	0.43
ANHU_16694	Arroyo Grande	San Luis Obispo	M	3124570	0.71
ANHU_16698	Arroyo Grande	San Luis Obispo	F	2685978	0.62
ANHU_16699	Arroyo Grande	San Luis Obispo	M	2316801	0.53
ANHU_16700	Arroyo Grande	San Luis Obispo	M	2350087	0.54
ANHU_16701	Arroyo Grande	San Luis Obispo	F	2793232	0.65
ANHU_16703	Arroyo Grande	San Luis Obispo	F	2125440	0.49
ANHU_16909	Goleta	Santa Barbara	F	2682251	0.62
ANHU_16910	Goleta	Santa Barbara	M	2165579	0.5
ANHU_16911	Goleta	Santa Barbara	F	2829951	0.65
ANHU_16912	Goleta	Santa Barbara	M	1976385	0.45
ANHU_16913	Goleta	Santa Barbara	M	1432311	0.33
ANHU_16914	Goleta	Santa Barbara	M	2551505	0.57
ANHU_16915	Goleta	Santa Barbara	M	1709372	0.39
ANHU_123	Winters	Yolo	M	11550861	2.66
ANHU_126	Winters	Yolo	F	20452527	4.63
ANHU_128	Winters	Yolo	M	11331387	2.52
ANHU_140	San Diego	San Diego	M	17170392	3.91
ANHU_160	San Diego	San Diego	F	22305221	5.07
ANHU_162	San Diego	San Diego	F	2746784	0.63
ANHU_168	San Diego	San Diego	F	1598805	0.37
ANHU_203	Beverly Hills	Los Angeles	F	16043405	3.69
ANHU_250	Beverly Hills	Los Angeles	M	17476632	3.94
ANHU_264	Beverly Hills	Los Angeles	F	18733004	4.29
ANHU_267	Beverly Hills	Los Angeles	F	12993554	3.02

ID	Town	County	Sex	Raw reads	Coverage
ANHU_323	Chico	Butte	F	16214478	3.7
ANHU_325	Chico	Butte	F	19315477	4.41
ANHU_334	Chico	Butte	M	21908223	4.99
ANHU_336	Chico	Butte	M	20660792	4.64
ANHU_338	Chico	Butte	F	19912775	4.47
ANHU_348	North Fork	Madera	M	22351753	5.1
ANHU_356	North Fork	Madera	F	23095567	5.31
ANHU_358	North Fork	Madera	F	25862170	5.68
ANHU_363	North Fork	Madera	M	19188387	4.4
ANHU_368	North Fork	Madera	F	16922702	3.93
ANHU_401	Winters	Yolo	F	8634737	1.98
ANHU_447	Grass Valley	Nevada	M	16563612	3.83
ANHU_450	Grass Valley	Nevada	M	22371483	5.11
ANHU_453	Grass Valley	Nevada	M	20993610	4.71
ANHU_467	Grass Valley	Nevada	F	21053303	4.77

Appendix B

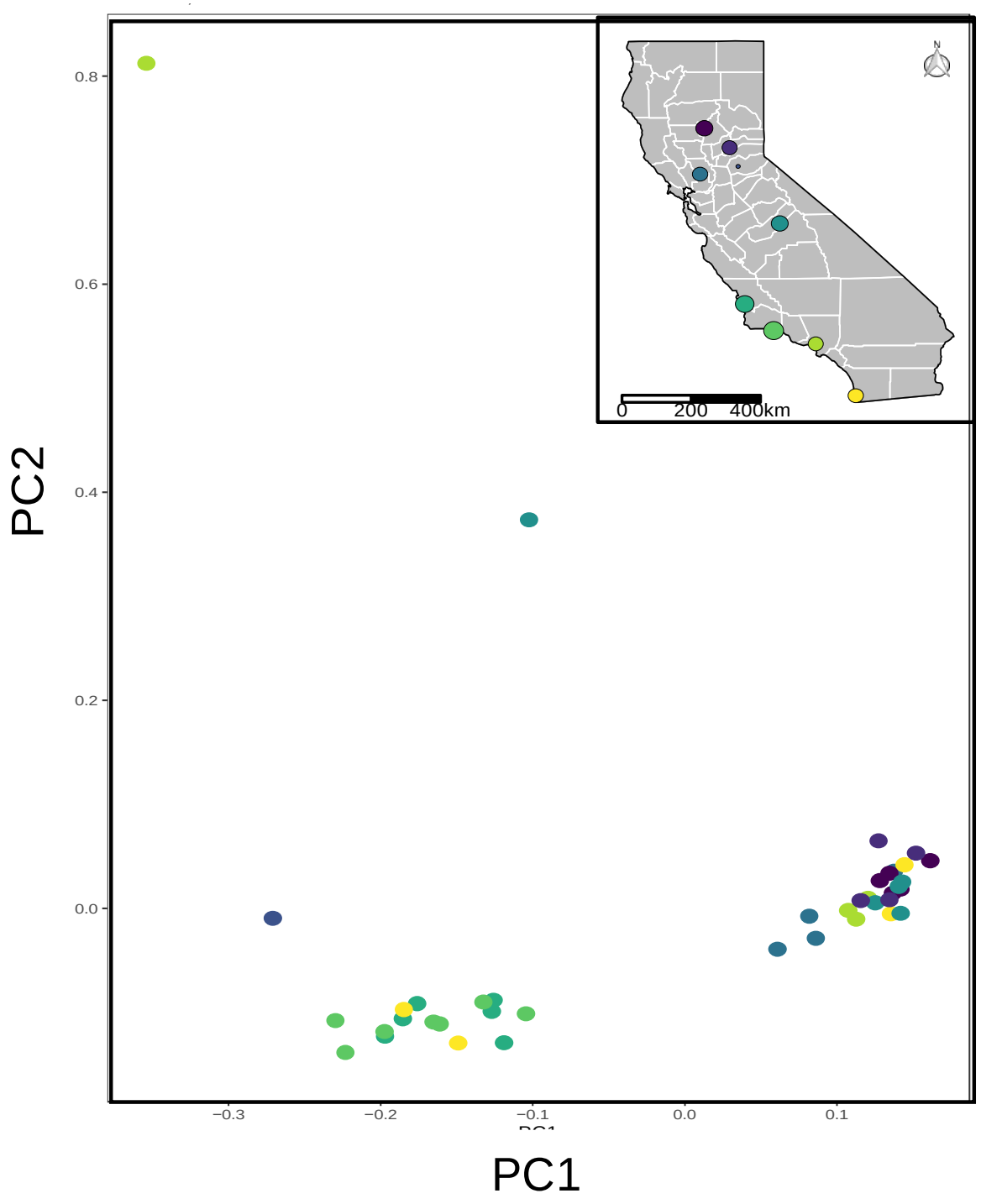

Figure A1. Principal components plot with all 40 individuals. Outlier individuals were removed to create plots within the main text.

Appendix C

Table A2. Position of SNPs used for testing of 100 SNP geographical assignment assay.

Chromosome	Position	F_{ST}
NW_022045445.1	27057	0.75505679
NC_044248.1	6446580	0.669777
NC_044264.1	10955874	0.42609921
NC_044264.1	10404878	0.42520199
NC_044264.1	10967229	0.41653195
NC_044264.1	10961031	0.41022547
NC_044264.1	10990660	0.40659707
NC_044264.1	10953697	0.40643103
NC_044244.1	139629753	0.3991544
NC_044248.1	8639379	0.3936843
NC_044264.1	10965117	0.38005465
NC_044246.1	5223334	0.37173349
NC_044254.1	29885599	0.36547495
NC_044245.1	108190239	0.36416597
NC_044253.1	24994040	0.36166471
NC_044245.1	100757564	0.3604218
NC_044260.1	12742533	0.36016052
NC_044246.1	7891338	0.35747038
NC_044255.1	1586942	0.3574533
NC_044244.1	195993881	0.3569053
NC_044247.1	5436972	0.35523923
NC_044244.1	149006997	0.3543643
NC_044254.1	28590866	0.34693881
NC_044253.1	30415806	0.34664271
NC_044244.1	157122285	0.34643659
NC_044264.1	10962538	0.34640778

Chromosome	Position	F_{ST}
NC_044264.1	10971850	0.34484038
NC_044254.1	25847386	0.34418975
NC_044245.1	22069694	0.34280446
NC_044244.1	141331468	0.34273198
NC_044253.1	20871530	0.33765315
NC_044244.1	2348359	0.33671448
NC_044244.1	85897371	0.33412635
NC_044257.1	18085719	0.33111532
NC_044271.1	5182369	0.3305189
NC_044248.1	19140767	0.33030803
NC_044244.1	40784302	0.31839274
NC_044248.1	28858451	0.31780281
NC_044245.1	37533099	0.31695383
NC_044246.1	108244182	0.31500895
NC_044244.1	115315128	0.31346223
NC_044273.1	3762188	0.31174187
NC_044264.1	10958303	0.31120164
NC_044248.1	35382670	0.3108084
NC_044245.1	105090385	0.30921795
NC_044244.1	56237646	0.30897768
NC_044276.1	24816831	0.30813632
NC_044253.1	7962289	0.30676025
NC_044253.1	27010223	0.30630275
NC_044248.1	39587224	0.30563319
NC_044248.1	14209549	0.30373946
NC_044244.1	55806595	0.30160756
NC_044274.1	12301018	0.30038308
NC_044244.1	163395873	0.30034364

Chromosome	Position	F_{ST}
NC_044265.1	751548	0.2973533
NC_044244.1	117931760	0.29589449
NC_044274.1	9590928	0.2954329
NC_044245.1	102440992	0.29383997
NC_044253.1	25444300	0.29321312
NC_044246.1	113921050	0.29312567
NW_022045449.1	89626	0.29075604
NC_044257.1	12935624	0.29068416
NC_044264.1	10984166	0.29004211
NC_044260.1	5720271	0.2895399
NC_044244.1	30674358	0.28932172
NC_044264.1	10951972	0.28847438
NC_044261.1	7848262	0.28763436
NC_044249.1	4022519	0.28749622
NC_044244.1	127540382	0.28590293
NC_044252.1	19970592	0.2857781
NC_044244.1	124954298	0.28512561
NC_044272.1	3584192	0.28262427
NC_044246.1	2215258	0.28220647
NC_044244.1	35162171	0.28192972
NC_044265.1	11326399	0.28173061
NC_044264.1	10994357	0.28166644
NC_044248.1	23119951	0.28137709
NC_044244.1	49423923	0.27788977
NC_044246.1	110436470	0.2777589
NC_044264.1	10985658	0.27731534
NC_044267.1	1503680	0.27670217
NC_044268.1	5249994	0.27666871

Chromosome	Position	F_{ST}
NC_044250.1	2196784	0.27604607
NC_044256.1	2620109	0.27490767
NC_044248.1	24027802	0.2732877
NC_044245.1	71569406	0.27316804
NC_044244.1	10910640	0.27284296
NC_044260.1	11266638	0.27263488
NC_044248.1	6777680	0.27250037
NW_022045535.1	36605	0.27234822
NC_044250.1	10664308	0.27156794
NC_044266.1	6048952	0.27107741
NC_044248.1	37584491	0.27106024
NC_044245.1	129266888	0.27049744
NC_044244.1	102998963	0.27043122
NC_044260.1	13694039	0.27041129
NC_044261.1	10887751	0.26965335
NC_044246.1	106801780	0.26937255
NC_044256.1	1485108	0.2675217
NC_044245.1	71450600	0.26729618

Appendix D

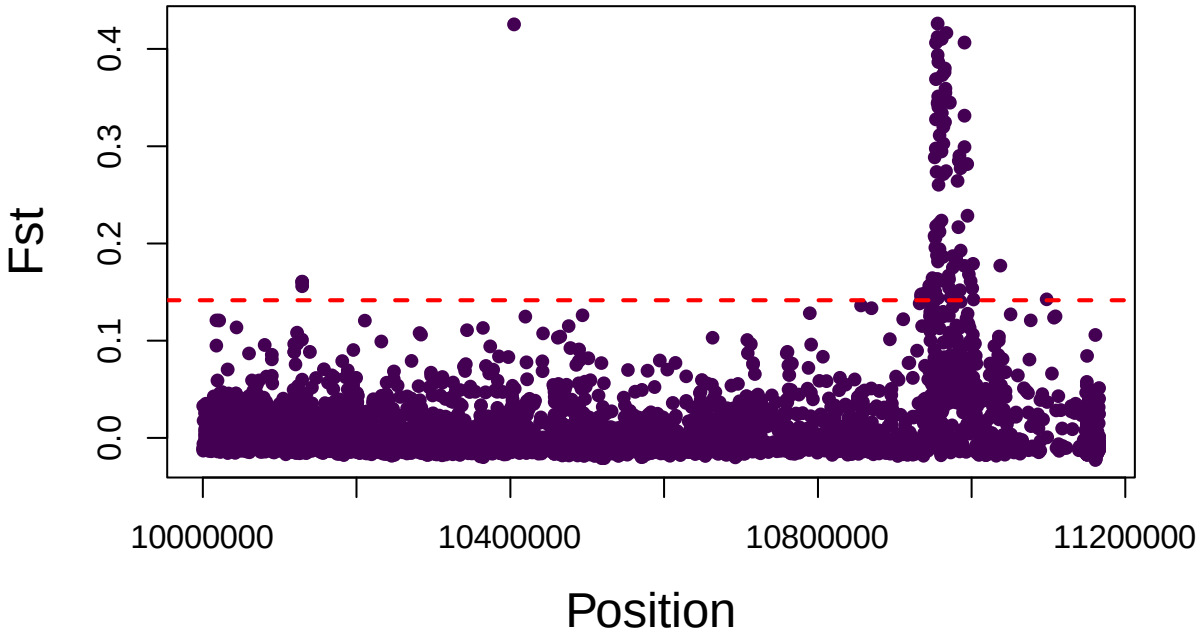

Figure A2. Manhattan plot showing F_{ST} peak between northern and southern populations. Such a region could be used as diagnostic locus predictive of breeding region.