# DIVERGENCE, GENE FLOW, AND THE SPECIATION CONTINUUM IN TRANS-BERINGIAN BIRDS

By

Jessica F. McLaughlin

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Science

in

Biological Sciences

University of Alaska Fairbanks

August 2017

APPROVED:

Kevin Winker, Committee Chair
Naoki Takebayashi, Committee Member
Kris Hundertmark, Committee Member
Kris Hundertmark, Chair
*Department of Biology and Wildlife*
Paul Layer, Dean
*College of Natural Science and Mathematics*
Michael Castellini, *Dean of the Graduate School*

## Abstract

Understanding the processes of divergence and speciation, particularly in the presence of gene flow, is key to understanding the generation of biodiversity. I investigated divergence and gene flow in nine lineages of birds with a trans-Beringian distribution, including pairs of populations, subspecies, and species, using loci containing ultraconserved elements (UCEs). I found that although these lineages spanned conditions from panmixia to fully biologically isolated species, they were not smoothly distributed across this continuum, but formed two discontinuous groups: relatively shallow splits with gene flow between Asian and North American populations, no fixed SNPs, and lower divergence; and relatively deeply split lineages with multiple fixed SNPs, higher divergence, and relatively low rates of gene flow. All eight lineages in which two populations were distinguishable shared the same divergence model, one with gene flow without a prolonged period of isolation. This was despite the diversity of lineages included that might not have responded in the same ways to the glacial-interglacial cycles of connection and isolation in Beringia. Together, these results highlight the role of gene flow in influencing divergence in these Beringian lineages.

Sample size is a critical aspect of study design in population genomics research, yet few empirical studies have examined the impacts of small sample sizes. Using split-migration models optimized with full datasets, I subsampled the datasets from Chapter 1 at sequentially smaller sample sizes from full datasets of 6 – 8 diploid individuals per population and then compared parameter estimates and their variances. Effective population size parameters (*v*) tended to be underestimated at low sample sizes (fewer than 3 diploid individuals per population), migration (*m*) was fairly reliably estimated until under 2 individuals per population, and no trend of over- or underestimation was found in either time since divergence (*T*) or $\Theta$ ($4N_{ref}\mu$) . Lineages that were split above the population level (subspecies and species pairs) tended to have lower variance at smaller sample sizes than population-level splits, with many parameters reliably estimated at levels as low as 3 diploid individuals per population, whereas shallower splits (i.e., populations) often required at least 5 individuals per population for reliable demographic inferences. Although divergence levels may be unknown at the outset of study design, my results provide a framework for planning appropriate sampling, and for interpreting results if smaller sample sizes must be used.

Table of Contents

## List of Figures

Page

## List of Tables

Page

# Acknowledgments

This study was supported in part by the National Science Foundation (DEB-1242267-1242241-1242260). The University of Washington Burke Museum provided some specimens for use in this study. We thank the National Center for Genome Analysis Support and the Research Computing Services of the University of Alaska Fairbanks. Oralee Nudson provided invaluable guidance on supercomputer use, and helpful bioinformatics advice was provided by Matthew Miller and Michael Harvey. We thank Kevin Hawkins, Phil Lavretsky, Jeff Peters, and Ryan Gutenkunst for help running δaδi. Thanks also to Kathryn Everson, Lindsey Klueber, Michael Saccone, and Alexandra Lewis for comments on earlier drafts of Chapter 1 (KE, LK, MS) and Chapter 2 (KE, AL). I would also like to thank Fern Spaulding, Rebecca Cheek, and Jack Withrow for their invaluable help, and to my family and friends for supporting me throughout my time at UAF.

## General Introduction

The history of Beringia is characterized by repeated cycles of connection and isolation between Asia and North America due to sea level changes between Pleistocene-era glacial and interglacial periods. These cycles could impact taxa whose ranges encompass both sides of the region, potentially allowing for periods of contact and gene flow between Asian and North American populations. This history may be reflected across various lineages of birds, and could shed light on the patterns and processes of divergence and speciation in the region.

Of particular interest is understanding the patterns of how divergence develops in Beringia, through which the processes driving divergence in Beringia can be inferred. I examined nine lineages of birds with a trans-Beringian distribution, including three pairs each of populations, subspecies, and species from three avian orders. These lineages span a range of phenotypic divergence from indistinguishable populations to well-differentiated, easily identifiable species. They also encompass diverse life histories, dispersal abilities, and potentially, histories of Beringian occupation, which might lead to different modes of divergence. Testing various models of how these lineages diverged might provide insights into how divergence occurred in this region. In Chapter 1, using best-fit models, I estimated both divergence ($F_{ST}$) and demographic parameters including gene flow, time since divergence, and effective population sizes to better understand the speciation continuum among Beringian birds.

For studies such as that in Chapter 1, care needs to be taken in study design. A key question is how many individuals are necessary for reliable demographic estimates. This is important for both the allocation of resources when determining the best plan for sampling populations, and for understanding how estimates might be impacted in scenarios where a study must proceed with a small sample size.

Although some previous work has used next-generation sequencing (NGS) data to investigate sample size impacts on population genetics estimates, none have used empirical data to specifically examine the demographic estimates of interest in my study. As the impact of low sample size is likely to vary among different estimates, it is vital to investigate this for specific parameters. My first chapter used sample sizes deemed optimal or above optimal for coalescent analyses and so provide excellent empirical datasets in which to investigate questions about the effects of smaller sample sizes. In Chapter 2, I resampled the individuals in the datasets used in Chapter 1 at iteratively smaller sample sizes to understand the impact of low sample sizes on estimates of effective population sizes, time since split, and gene flow.

# Chapter 1: Divergence, gene flow, and the speciation continuum in nine lineages of trans-Beringian birds[1]

## 1.1 Abstract

Understanding the processes of divergence and speciation, particularly in the presence of gene flow, is key to understanding the generation of biodiversity. We investigated divergence and gene flow in nine lineages of birds with a trans-Beringian distribution, including pairs of populations, subspecies, and species, using loci containing ultraconserved elements (UCEs). We found that although these lineages spanned conditions from panmixia to fully isolated species, they were not smoothly distributed across this continuum, but formed two discontinuous groups: relatively shallow splits with gene flow between Asian and North American populations, no fixed SNPs, and lower divergence; and relatively deeply split lineages with multiple fixed SNPs, higher divergence, and relatively low rates of gene flow. All eight lineages in which two populations were distinguishable shared the same divergence model, one with gene flow without a prolonged period of isolation. This was despite the diversity of lineages included that might not have responded in the same ways to the glacial-interglacial cycles of connection and isolation in Beringia. Despite correlation between our UCE-based estimates of gene flow and previous estimates of divergence from mitochondrial DNA, we found that discord prevails among divergence estimates from mtDNA, AFLPs, and UCEs. Together, these results highlight the role of gene flow in influencing divergence in these Beringian lineages and of the presence in this region of discontinuous patterns of divergence.

[1] McLaughlin JF, Faircloth BC, Glenn TC, Winker K. Divergence, gene flow, and the speciation continuum in nine lineages of trans-Beringian birds. Prepared for submission in *Molecular Ecology*.

## 1.2 Introduction

The study of the processes of population divergence and speciation is key to understanding the generation of biodiversity. The genomic basis of divergence, though, is influenced by multiple mechanisms, including selection, mutation, and drift. However, gene flow also plays an important role, as migration can counteract differentiation due to selection and/or drift when it occurs during primary divergence or secondary contact (Price 2008, Seehausen *et al.* 2014), leaving recognizable signatures across the genome that differ from divergence between allopatric populations (Sousa & Hey 2013). Understanding gene flow, therefore, is vital to understanding the patterns of divergence and speciation in populations that may not be strictly isolated.

Of particular interest is determining whether there are broad patterns in how divergence develops that can provide insight into the mechanisms of speciation. Are there common patterns in how populations diverge, particularly within a shared geographic system, that can reveal how the divergence-to-speciation continuum develops, and how this process plays out across different parts of the genome? It is well known that various portions of the genome will have different levels of divergence, depending on a variety of factors including, among others, differing inheritance patterns (Funk & Omland 2003, Avise 2004, Toews & Brelsford 2012), linkage with genes undergoing natural or sexual selection (Via & West 2008, Feder *et al.* 2012, Casillas & Barbadilla 2017, Wolf & Ellegren 2017), the structure and arrangement of the genome itself (Delmore *et al.* 2015, Ragland *et al.* 2017, Vijay *et al.* 2017, Wolf & Ellegren 2017), and the history of demographic events in the populations (Sousa & Hey 2013, Casillas & Barbadilla 2017). Measures of divergence from various markers are therefore frequently discordant due to this heterogeneity (Humphries and Winker 2011, Peters *et al.* 2014), and inference of the extent and history of gene flow may also be skewed by reliance on a single marker type (Cahill *et al.* 2015, Good *et al.* 2015, Zarza *et al.* 2016).

Species- and subspecies-level taxonomic assignments, meanwhile, break a continuous process into discrete bins that tend to be heavily weighted toward phenotype, so taxonomy may not reflect levels of divergence. Studies in multiple systems have found that striking phenotypic differences can result from a few changes in relatively small genomic regions, with ongoing gene flow between distinct phenotypic forms (e.g., Toews *et al.* 2016, Van Belleghem *et al.* 2017). Therefore, genomic estimates of divergence cannot be assumed to be correlated with, and may in fact be discordant with, taxonomic assignment based on such phenotypic variation, and may provide more precise estimates of where diverging populations are in the process of speciation. However, as divergence will be heterogeneous across the genome, individually each type of genetic marker may produce estimates of gene flow and divergence that are at odds with other marker types (Nosil *et al.* 2009, Humphries & Winker 2011, Ellegren *et al.* 2012). When considered across multiple taxa at different points on the continuum of speciation, patterns in not only gene flow and divergence, but also in the discord between these estimates between different genetic markers, may emerge that illuminate the role of specific mechanisms, especially how gene flow is involved with the dynamics of the speciation process.

One strategy for understanding the patterns of divergence and speciation, and thus the underlying processes, is to make comparisons among multiple species within a common geographic framework that occupy different points along the process of speciation, from a single panmictic population to fully reproductively isolated populations. This approach is most effective if a large number of orthologous markers are used, because this would increase the portion of the genome included in analyses while also allowing direct comparison of parameters between different lineages (Harvey *et al*. 2016). One type of marker that may be well suited to this approach is loci containing ultraconserved elements (UCEs), which are centered on highly conserved genomic regions (≥ 80% identity across ≥ 100 bp) distributed throughout animal genomes (Bejerano *et al*. 2004, Siepel *et al.* 2005, Stephen *et al.* 2008, Faircloth *et al.* 2012). The highly conserved core allows for orthologous loci to be sequenced across a wide range of taxa, while the flanking regions accumulate mutations with increasing distance from the core (Faircloth *et al.* 2012), enabling inferences of population history even at relatively shallow levels of divergence.

We used loci with UCEs to examine divergence and gene flow in nine lineages of trans-Beringian birds to better understand the process of avian speciation in a region in which gene flow has likely been a factor. These same lineages were examined previously with mitochondrial and amplified fragment length polymorphism (AFLP) data, enabling us to make comparisons among patterns of divergence in multiple datasets (Humphries & Winker 2011). Discord in estimates of divergence from mitochondrial and nuclear markers has previously been found in some of these lineages (Humphries & Winker 2011), indicating heterogeneity in the genomic landscape of divergence. Furthermore, these estimates were not correlated with phenotypic divergence as indicated by taxonomy. We compared our UCE results to these earlier results to build a more comprehensive picture of patterns of divergence in Beringia. Overall, we ask how divergence and speciation have developed among avian lineages across a shared geographic region, including whether there are any common patterns in demographic history among lineages, such as in modes of divergence (speciation with gene flow, strict isolation, isolation followed by secondary contact), and where each lineage is along the speciation continuum.

## 1.3 Methods

### *1.3.1 Study system*

Through the Pleistocene (2.6 million - 10,000 years ago), central Beringia experienced multiple cycles of exposure and inundation from sea level changes driven by glacial cycles, resulting in multiple periods of connectivity between Eurasia and North America, followed by isolation (Hopkins 1959, Hopkins *et al.* 1965). This episodic cycling, which is estimated to have occurred at least nine and possibly up to twenty times over the course of the Pleistocene (Hopkins 1967, Pielou 1991), potentially would have repeatedly connected the biota of Eurasia and North America, followed by isolation. This pattern has the potential to be reflected in various estimates of timing and degree of divergence between lineages occurring on both sides of Beringia, as different population pairs may have split during different flooding events throughout the past 2.6 million years.

We examined nine lineages of birds, each representing a pair of populations, subspecies, or species in each of three orders— Asian and North American populations of *Clangula hyemalis* (long-tailed duck), *Anas crecca crecca/Anas crecca carolinensis* (green-winged teal), and *Anas penelope/Anas americana* (Eurasian and American wigeons) in Anseriformes; *Pluvialis squatarola* (black-bellied plover), *Numenius phaeopus variegatus/Numenius phaeopus hudsonicus* (whimbrel), and *Tringa brevipes/Tringa incana* (gray-tailed and wandering tattlers) in Charadriiformes; and *Luscinia svecica* (bluethroat), *Pinicola enucleator kamschatkensis/Pinicola enucleator flammula* (pine grosbeak), and *Pica pica/Pica hudsonia* (Eurasian and black-billed magpies) in Passeriformes. Although we anticipated based on previous work (Humphries & Winker 2011) that these taxonomic levels of population, subspecies, and species were not well correlated with genetic divergence between the populations in each lineage, these classifications do reflect levels of phenotypic divergence, which is useful to examine relative to our genomic data.

Following Humphries and Winker (2011), we used current taxonomic assignment as a surrogate for phenotypic divergence.

### *1.3.2 Laboratory procedures*

Archived museum specimens were sampled for all taxa (Table S1).We aimed to extract whole genomic DNA from 8 individuals per population in each lineage, but fewer specimens were available in some groups, so smaller sample sizes of 5-7 individuals per population were used in those cases. Double-indexed DNA libraries were then prepared for each sample as described in Glenn *et al.* (2016), which were then each quantified with a Qubit fluorimeter (Invitrogen, Inc.). The libraries were then combined into equimolar pools. The pools were enriched using the Tetrapods-UCE-5Kv1 kit (MYcroarray) for a set of 5,060 loci, using UCE enrichment protocol version 1.5 and post-enrichment amplification protocol version 2.4 (ultraconserved.org) with HiFi HotStart polymerase (Kapa Biosystems) and 14 cycles of post-enrichment PCR. Distribution of fragment size of the enriched pool was quantified on a BioAnalyzer (Agilent, Inc), and the enriched pool was quantified by qPCR with a commercial kit (Kapa Biosystems). We sequenced all pools using an Illumina HiSeq 2500.

### *1.3.3 Bioinformatics*

After sequencing, data were demultiplexed with bcl2fastq (v 1.8.4; Illumina, Inc.) and adapters and low-quality bases trimmed using Illumiprocessor (Faircloth 2013), a parallel wrapper around Trimmomatic (Bolger *et al.* 2014). The singleton and read 1 fastq files for each individual were combined and then, with read 2 files, were assembled with Trinity (v2.0.6; Grabherr *et al*. 2011) run on Galaxy (Afgan *et al.* 2016). UCE loci were extracted and a complete matrix was constructed for each lineage using PHYLUCE (v 1.5; Faircloth 2016), providing information on the median number of loci shared and unshared by individuals in each lineage's dataset.

In each population pair, the two individuals in each population with closest to the median number of loci were identified, and the fastq sequence files for the four individuals were combined to produce a single read 1 and read 2 file for each. This was done to build a reference against which to call variants for all individuals, reasoning that such a middle-quality reference given the entire dataset balances retention and loss of loci due to quality control issues farther along the pipeline (reference sequence datasets are archived as supplemental files; a list of these files can be found in the General Appendix). These sequences were then assembled with Trinity on Galaxy as above, and PHYLUCE was used to create a fasta file of UCE loci. This was then indexed with BWA and SAMtools (Li & Durbin 2009; Li *et al.* 2009) for calling single nucleotide polymorphisms (SNPs).

SNPs were called using a modified workflow for population genomics with UCEs developed by Faircloth and Michael Harvey (https://github.com/mgharvey/seqcap_pop). For each dataset, sequences were aligned to the reference with BWA-MEM (Li 2013), and the resulting SAM alignments were converted to BAM with SAMtools. Alignments were checked for BAM format violations, read-groups header information was added, and PCR duplicates were marked for each individual using Picard (http://broadinstitute.github.io/picard). The resulting BAM files for each individual were merged into a single file with Picard, which was then indexed with SAMtools. The Genome Analysis Toolkit (GATK; v 3.4-0; McKenna *et al*. 2010) was then used to locate and realign around indels, which was followed by calling SNPs using the UnifiedGenotyper tool in GATK. SNPs and indels were then annotated and indels masked. We then restricted our datasets to high-quality SNPs (Q30) and performed read-backed phasing. We added additional filters with VCFtools (Danecek *et al.* 2011), reducing our dataset to a complete matrix with a minimum genotype quality (GQ) of 10. To confirm that invariant loci were retained due to quality, rather than missing data, we used the GATK function EMIT_ALL_CONFIDENT_SITES, followed by filtering to remove loci with inadequate data.

### *1.3.4 Analyses*

VCFtools was used to calculate coverage depths and both SNP- and locus-specific $F_{ST}$. The datasets were converted to STRUCTURE format (Falush *et al.* 2003) with PGDSpider (v 2.0.9.1; Lischer & Excoffier 2012). We then used the R package adegenet (v 3.2.2, R Core Team 2015; Jombart and Ahmed 2011) to test for Hardy-Weinberg equilibrium in each population, and calculated observed and expected heterozygosity and $F_{ST}$ (using the *G*-test with 99 bootstraps in all lineages except *Pluvialis squatarola*, in which 10,000 bootstraps were used), and assignment probabilities using DAPC (Discriminant Analysis of Principal Components).

We used diffusion analysis for demographic inference (δaδi; Gutenkunst *et al.* 2009) to estimate parameters of effective population size for both North American and Asian populations ($v_{NA}$ and $v_A$, respectively), migration ($m$), time since split ($T$), and $\Theta$, defined as $4N_{ref}\mu$, with $N_{ref}$ defined as ancestral population size and $\mu$ as substitution rate per generation, from which biologically meaningful values of effective population size ($N_{NA}$ and $N_A$), migration per generation ($M$), time since split in years ($t$), and ancestral population size ($N_{ref}$) were calculated. To prepare our data for δaδi, the phased SNPs were thinned with VCFtools to 1 biallelic SNP per 2000 base pairs. As most loci were well below this length (average locus length ranged between 673 - 1232 bp), this selected the first SNP on each locus. We then used a custom script (https://github.com/jfmclaughlin92/thesis) to remove Z-linked loci by identifying loci aligned with high probability by BLASTn (Zhang *et al.* 2000) to the Z chromosome of *Gallus gallus* (for Anseriformes and Charadriiformes; NCBI Annotation Release 103) or *Taeniopygia guttata* (for Passeriformes; NCBI Annotation Release 103) and removing them from the vcf file. The remaining SNP data were then converted into the joint site frequency spectrum (SFS) format using a perl script by Kun Wang (https://groups.google.com/forum/#!msg/dadi-user/p1WvTKRI9_0/1yQtcKqamPcJ). We then ran two-population models using δaδi, except in *Pluvialis squatarola*, which did not show significant population structure between Asian and North American populations.

For those datasets that did show significant genetic structure between the two populations, as estimated with adegenet, four different two-population models were run: neutral, isolation with migration (“IM”), isolation without migration (“island”), and split migration (“splitmig”; Figure 1.1). Upper and lower bounds were optimized by running each model repeatedly until the highest maximum log composite likelihood value was observed consistently between multiple runs with the same parameter bounds without parameter estimates pushing the bounds. After the best model was determined for each dataset, we ran it with 100 bootstrapped datasets (constructed with a Python script that resampled individuals with replacement; https://github.com/jfmclaughlin92/thesis) to estimate the 95% confidence intervals (CI) for each parameter.

In the case of *Pluvialis squatarola*, the whole dataset was treated as one population and a series of one-population δaδi models were run to estimate population growth and contraction over time: neutral, instantaneous population growth after a time point (“two epoch”), exponential growth after a time point (“expgrowth”), a population contraction and expansion (“bottlegrowth”), and a population contraction and expansion with specified beginning and ending time points (“three epoch”). Model optimization was carried out as above, and the best-fit model (three epoch) bootstrapped 100 times (https://github.com/jfmclaughlin92/thesis).

To interpret the model parameter estimates in biological terms, we first BLASTed each reference fasta against the NCBI-available genome with the closest relationship to the lineage in the dataset, using the time since most recent common ancestor (TMRCA) estimates from Claramunt and Cracraft (2015), as summarized in Table S2. After importing BLAST results, we removed lower-affinity duplicate hits and then tallied total base pairs (bp) and substitutions, which were used to calculate substitutions per site. This was then converted to substitutions per site per year by multiplying by TMRCA*2. Generation time ($G$) was determined as $G = \alpha + (s/(1 - s))$, where $\alpha$ is the age at first breeding and $s$ is annual adult survival, following the method in Saether *et al.* (2005) (Table S2). This was then used to convert the yearly substitution rate to substitutions per site per generation.

To investigate the relationships between parameter estimates, divergence, and taxonomic assignment, we tested for correlations of $T$ and $m$ obtained from our UCE datasets using δaδi with $F_{ST}$ as measured from UCEs, mtDNA, and AFLPs (the latter two from Humphries & Winker 2011); $D_A$ (Humphries & Winker 2011); and current taxonomic placement (using 0 to represent population pairs, 0.5 for subspecies, and 1 for species).

### 1.4 Results

We obtained over 200 million reads (sequence fragments), with reads per specimen ranging between 379,344 and 4,010,381, with an average of 1,450,760, of which > 99% passed adapter and quality control trimming. Assembly of the reference produced between 130,506 and 657,330 contigs, totaling 47,215,417 – 254,336,867 bp. All datasets produced more than 1,000 loci over 1 Kb (range 1,086 – 9,935; Table S3). We identified 4,040-4,294 UCE loci in each reference dataset, with average contig length between 673 and 1,232 bp (Table S4). An average of 54.2 % of loci were variable, and an average of 1.99 – 7.42 SNPs per locus were called in these loci. In total, 3,254 – 13,215 SNPs were called in each dataset, and thinning to one SNP per locus left 1,636 – 2,656 SNPs (Table 1.1). Coverage across all SNPs averaged 35.1X, ranging between 30.4 X – 38.6 X (Table 1.1). Expected heterozygosity ($H_E$) ranged from 0.079 to 0.160, and observed heterozygosity ($H_O$) between 0.086 and 0.179 (Table 1.2). In three lineages (*Numenius phaeopus*, *T. brevipes/T. incana*, and *Pica pica/Pica hudsonia*), there were significant differences after correcting for multiple tests between $H_E$ and $H_O$ and between Asian and North American populations (Table 1.2).

In eight of the nine lineage pairs, significant positive $F_{ST}$ values between Asian and North American populations were obtained, ranging between 0.004 and 0.58 (Table 1.3). Four lineages (*Anas penelope/A. americana*, *A. c. crecca/A. c. carolinensis*, *C. hyemalis*, and *L. svecica*) had overall between-population $F_{ST}$ values below 0.05, whereas the four other lineages all had $F_{ST}$ values above 0.2. In the lower-$F_{ST}$ group, there were no fixed SNPs in the full datasets (all SNPs, including Z-linked loci); in the higher-$F_{ST}$ group, the number of fixed SNPs ranged from 12 – 121 in the one-SNP-per-locus datasets and between 31 and 299 in the full datasets (Table 1.3).

For all two-population datasets, the split-migration model provided the best fit (Table 1.4). With these models, $N_{ref}$ (ancestral population size) was estimated at between 4,408 (*Tringa*) and 37,561 (*Pica*) individuals (Table 1.5). Estimates of effective population sizes in Asian populations ranged between 19,867 (*Tringa brevipes*) and 207,593 (*Luscinia svecica*), whereas those of North American populations were between 8,434 (*Tringa incana*) and 138,918 (*Luscinia svecica*). In three lineages, the Asian population was markedly larger than the North American population: tattlers (19,867 vs 8,434), magpies (128,078 vs 54,899), and whimbrel (22,661 vs 12,368). Estimates of time since divergence ranged between 153,995 and 364,903 yr, with a mean of 241,378 yr. Overall, estimates of *m* varied between 0.005 and 0.574, with calculations of migrants-per-generation (*M*) between 0.003 and 1.440 (Table 1.5). However, all of the lower-$F_{ST}$ lineages had migration rates of more than 0.684 birds/generation, whereas all higher-$F_{ST}$ lineages had low migration rates (less than 0.014 birds/generation).

The best-fit single-population model for the *Pluvialis squatarola* dataset was found to be the "three-epoch" model of growth following a population contraction and expansion (Table 1.6). We found that the Beringian population of this species showed evidence of a population contraction and expansion 1.8 Mya, in which their effective population fell from 73,413 (± 7,187) to 10,077 (± 1,059) individuals, then rebounded to 225,682 (± 15,158; Table 1.6).

There was a significant correlation between estimates of *m* and $F_{ST}$ from UCE datasets (Adj $R^2$ = 0.6686, *p* = 0.0081, Figure 1.2). There was an additional correlation between $\nu_{NA}$ (effective North American population size) and $F_{ST}$ (Adj $R^2$ = 0.5764, *p* = 0.017). However, other parameter estimates were not correlated with $F_{ST}$ estimates from UCEs, including *T* (time since divergence), *Θ*, and $\nu_A$ (effective Asian population size). We also examined correlations between marker classes, finding no significant relationships between any of the estimates of $F_{ST}$ from UCEs, mtDNA, and AFLPS. In general, the highest divergence estimates were found in mtDNA, followed by UCEs, then AFLPs, with most lineages (with the exception of *Pluvialis squatarola*) that had insignificant estimates of $F_{ST}$ in mtDNA and/or AFLPs having significant $F_{ST}$ from UCE data. Among marker classes there was a relationship between mitochondrial divergence ($F_{ST}$) and *m* (migration rate) estimated from UCEs (Adj $R^2$ = 0.5932, *p* = 0.0154), and between current taxonomic assignment and *m* (Adj $R^2$ = 0.5145, *p* = 0.02728). Other demographic parameters ($\nu_A$, *T*, *Θ*) did not correlate with divergence estimates from AFLPs or mitochondrial DNA.

## 1.5 Discussion

For all eight two-population lineages, a single model type (split-migration; Figure 1.1D) provided the best fit for the data. This suggests that speciation with gene flow is the predominant mode of speciation among birds in this region. This similarity among these lineages occurs despite the fact that they span three avian orders and have diverse life histories, seasonal migration behaviors, dispersal abilities, habitat requirements, and, possibly, Beringian occupation times. Each of these could influence how each lineage responded to glacial/interglacial cycles of connection and isolation. Habitat availability would have varied considerably across Beringia spatially and between glacial-interglacial cycles throughout the Pleistocene, depending on climatic conditions (Melles *et al.* 2012), creating a mosaic of habitat types across the past 2.6 million years that would have provided increased opportunities for population connectivity in some lineages, likely influencing variations in the levels of gene flow observed. For groups that would not disperse as well across habitat types that were not favorable, however, a strong signature of long-term isolation (leading to an IM model being a better fit) might be expected. Yet in all eight lineages, the best-fit model was the same—a split migration model with variable levels of gene flow, suggesting speciation-with-gene-flow rather than long periods of isolation followed by secondary contact as the dominant process of avian divergence in Beringia.

These lineages in Beringia encompassed a range of levels of divergence, from a continuous population spanning both the Asian and North American sides of the Bering Strait (*Pluvialis squatarola*) to well-diverged populations with low amounts of gene flow (*T. brevipes/T. incana, N. phaeopus variegatus/hudsonicus, Pica pica/Pica hudsonia, Pinicola enucleator kamschatkensis/flammula*). However, these were not distributed across a smooth divergence continuum, but seemed to cluster into two broad groups: a single population ($F_{ST}$ = 0.0) and lower-divergence group ($F_{ST}$ = 0.004 - 0.044; mean $M$ = 0.98 individuals/generation); and a higher-divergence group ($F_{ST}$ = 0.269 - 0.585), with sharply decreased levels of gene flow (mean $M$ = 0.022 individuals/generation). These groups did not correspond with the three taxonomic levels represented in the study, nor with the estimates of $F_{ST}$ from other marker types, highlighting the heterogeneous nature of divergence (Figure 1.2).

Previous work (Hendry *et al.* 2009, Flaxman *et al.* 2014, Roux *et al.* 2016, Nosil *et al.* 2017, Riesch *et al.* 2017) has suggested that the speciation process can be a two-state system, with few populations existing in the middle ground between near to complete panmixia and fully reproductively isolated species. Furthermore, periods of gene flow can promote the formation of such a dynamic (Flaxman *et al.* 2014, Nosil *et al.* 2017, Riesch *et al*. 2017). When gene flow occurs, the feedback process of divergent selection and linkage disequilibrium on the background of genomic architecture can cause populations that have begun to diverge and come into contact following isolation to return to a single well-mixed population unless a critical level of differentiation has already been achieved (Flaxman *et al.* 2014). Given the cyclical nature of population isolation and connectivity in Beringia, this bimodal pattern may be more likely to develop here, rather than favoring the development of stable middle states (Flaxman *et al*. 2014), and different parts of the genome will be impacted by this process at different rates. Our data are concordant with the model of two steady states of divergence, with no taxa observed in the intermediate region of gene flow or genetic divergence.

Future studies of speciation in Beringia could examine the effects of selection and drift, and should seek to bridge two sampling gaps. First, sampling should take place encompassing more lineages to better understand the nature of the speciation continuum, particularly whether there is in fact a two-phase dynamic present. Secondly, a larger portion of the genome should be sampled to clarify the precise role of gene flow relative to other factors, particularly selection and drift, in impacting this process. Although UCEs are useful for the analyses used here, we cannot use them for a detailed examination of the role of divergent selection in preventing reticulation in lineages with potential gene flow, and whether there are any trends in the strength of selection relative to levels of gene flow. This is particularly important because gene flow and selection are tightly linked in divergence with gene flow, with ever greater selection needed to overcome increasing amounts of gene flow if speciation is to proceed (Coyne & Orr 2004, Price 2008, Sousa & Hey 2013, Seehausen *et al.* 2014).Together, these would allow us to improve our understanding of how avian divergence and speciation in Beringia have been influenced by the history of isolation and connection between Asia and North America.

Earlier work in these lineages found a remarkable degree of discordance between nuclear and mitochondrial estimates of divergence (Humphries & Winker 2011). As a new marker class for population genomics, we did not expect UCEs to resolve such discord; instead, they add to it, lacking significant correlation with divergence estimates from both AFLPs and mtDNA (Figure 1.3). Some of this discord is likely due to the different effective population sizes between the marker types, with mtDNA having the smallest $N_e$ and highest divergence, UCEs being intermediate, and AFLPs the highest $N_e$ and lowest $F_{ST}$. Although our estimates of divergence from UCEs were concordant with those from mtDNA and AFLPs in some lineages when these patterns were taken into account (*T. brevipes/T. incana* and *Pica pica/Pica hudsonia* having relatively high estimates for each type of marker, and *Pluvialis squatarola* consistently not found to have significant divergence), we found a small but significant level of divergence in both *C. hyemalis* and *L. svecica* that had not been previously found (Figure 1.3). Additionally, some lineages that had been previously found to be discordant between mtDNA and AFLPs had significant UCE-based $F_{ST}$ estimates despite insignificant estimates from AFLPs (*Numenius phaeopus*, *Pinicola enucleator*). This reinforces the hypothesis that a strong degree of heterogeneity in divergence between different parts of the genome exists during divergence and speciation.

Some of our estimates of gene flow differed from levels found in other multilocus nuclear studies, for example being considerably lower between *A. penelope* and *A. americana* than previously found (Peters *et al.* 2014). Additionally, our estimates of gene flow between subspecies of *A. crecca*, with an estimated 0.70 – 0.84 individuals per generation in both directions, were markedly different than the strong asymmetry found in previous studies, which found no detectable gene flow into *A. c. crecca* but much higher (20-28 individuals per generation) into *A. c. carolinensis* (Peters *et al.* 2012, Winker *et al.* 2013). Current taxonomy also does not reflect the genomic patterns found here. In particular, *Numenius phaeopus* subspecies have the opportunity for contemporary gene flow, yet have near-zero gene flow, whereas *A. penelope* and *A. americana* have higher rates than would be expected in species that are effectively reproductively isolated (see also Peters *et al.* 2014). Taxonomic revisions may be warranted in these cases.

The divergence and speciation processes in these Beringian birds are best summarized by a single model framework of speciation with gene flow. Gene flow is thus an important factor in the generation and maintenance of biodiversity in this system, and it was the only factor that correlated with divergence ($F_{ST}$). The two clusters of $F_{ST}$ estimates were reflected by two clusters of gene flow estimates, and no signature of long-term isolation followed by secondary contact was found.

### 1.6 Figures

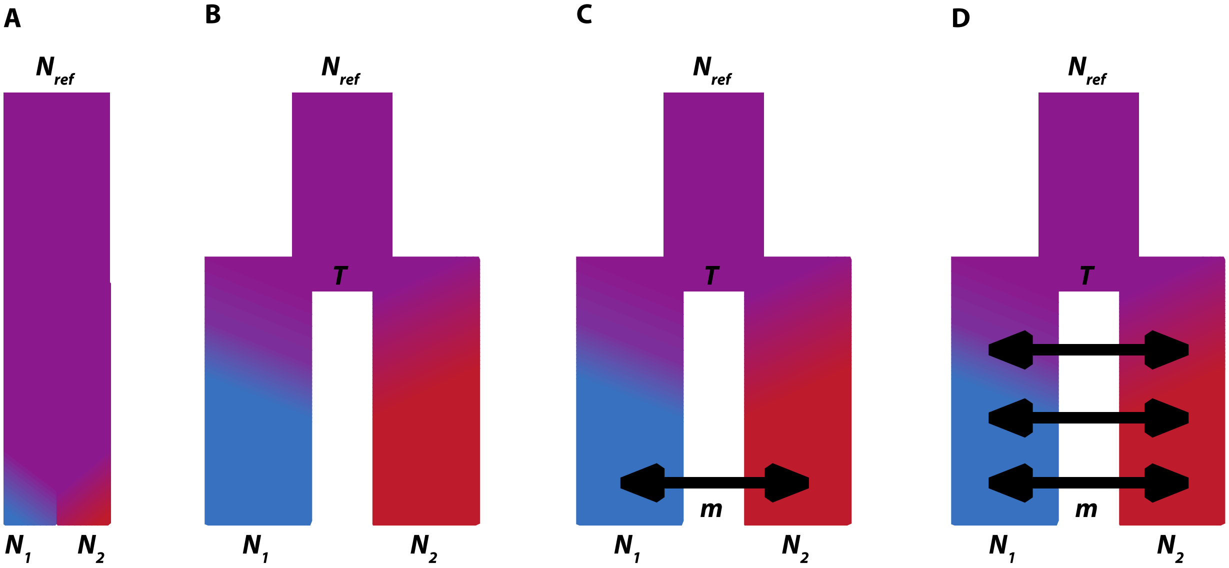

Figure 1.1: Models of divergence tested with δaδi on two-population splits: (A) neutral model, (B) isolation without migration ("island"), (C) isolation with migration ("IM"), and (D) split migration, speciation with gene flow ("splitmig").

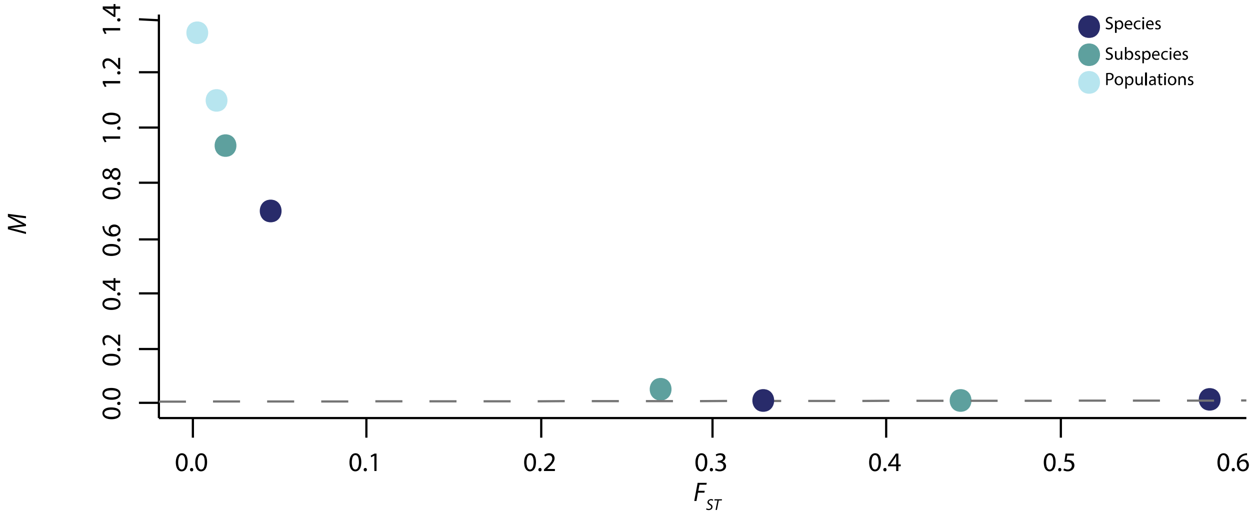

Figure 1.2: UCE-based estimates of $F_{ST}$ vs migration rate in individuals per generation ($M$) in the eight lineages in which two-population models were appropriate. Color indicates taxonomic level of each pairwise comparison.

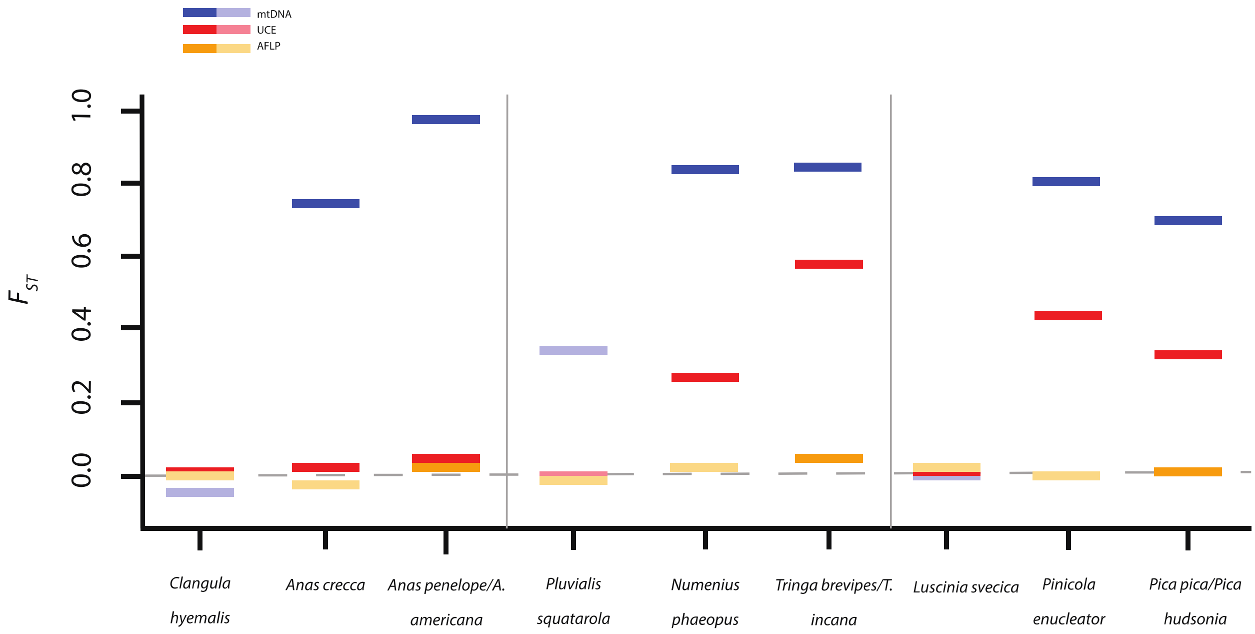

Figure 1.3: $F_{ST}$ comparisons between this study and Humphries and Winker (2011) for nine lineages in three orders (orders separated by vertical gray lines), with lineages given in the sequence population-subspecies-species in each order. Bold tones indicate significant $F_{ST}$ estimates, while pale tones indicate insignificant ones. Estimates below zero are due to computational idiosyncrasies and are effectively zero.

**1.7 Tables**

Table 1.1: Summary of single nucleotide polymorphisms (SNPs), including number of individuals in each lineage, total number of SNPs, average coverage per SNP, number of SNPs after thinning to one SNP per locus, and average number of SNPs per locus (variable loci only)

| | **N (Asia/North America)** | **Total SNPs** | **Average coverage depth per SNP** | **SNPs per locus (variable loci only)** | **Thinned SNPs (total variable loci)** |
|---|---|---|---|---|---|
| **Anseriformes** | | | | | |
| *Clangula hyemalis* | 8:8 | 9,276 | 30.76 | 3.798 | 2,442 |
| *Anas crecca* | 8:6 | 9,022 | 38.64 | 3.636 | 2,481 |
| *Anas penelope /A. americana* | 8:8 | 7,041 | 37.61 | 3.041 | 2,315 |
| **Charadriiformes** | | | | | |
| *Pluvialis squatarola* | 5:5 | 13,215 | 30.41 | 7.420 | 1,781 |
| *Numenius phaeopus* | 8:7 | 6,492 | 31.15 | 2.718 | 2,388 |
| *Tringa brevipes /T. incana* | 8:8 | 3,254 | 37.99 | 1.989 | 1,636 |
| **Passeriformes** | | | | | |
| *Luscinia svecica* | 8:7 | 9,379 | 38.06 | 3.728 | 2,516 |
| *Pinicola enucleator* | 8:7 | 8,117 | 36.48 | 3.056 | 2,656 |
| *Pica pica/Pica hudsonia* | 7:7 | 9,276 | 34.90 | 4.218 | 2,199 |

Table 1.2: Expected ($H_E$) and observed heterozygosities ($H_O$) for each population. Asterisks (*) indicate significant differences after false discovery rate correction.

| | **$H_E$ (Asia/North America)** | ***p* (Asia vs NA)** | **$H_O$ (Asia/North America)** | ***t*** | ***df*** | ***p* ($H_E$ vs $H_O$)** |
|---|---|---|---|---|---|---|
| **Anseriformes** | | | | | | |
| *Clangula hyemalis* | 0.118/0.116 | 0.856 | 0.1723/0.1790 | -9.352 | 2441 | 1 |
| *Anas crecca* | 0.112/0.113 | 0.576 | 0.1203/0.1246 | -2.486 | 2323 | 0.994 |
| *Anas penelope /A. americana* | 0.105/0.111 | 0.03 | 0.1144/0.1235 | -3.935 | 2215 | 1 |
| **Charadriiformes** | | | | | | |
| *Pluvialis squatarola* | 0.155/0.160 | 0.68 | 0.1723/0.1790 | -6.687 | 1780 | 1 |
| *Numenius phaeopus* | 0.134/0.108 | 0.002 * | 0.1434/0.1210 | 9.498 | 2387 | $2.2 \times 10^{-16}$ * |
| *Tringa brevipes /T. incana* | 0.119/0.079 | 0.002 * | 0.1319/0.0856 | 16.201 | 1566 | $2.2 \times 10^{-16}$ * |
| **Passeriformes** | | | | | | |
| *Luscinia svecica* | 0.134/0.128 | 0.078 | 0.1471/0.1419 | -6.937 | 2515 | 1 |
| *Pinicola enucleator* | 0.108/0.142 | 0.034 | 0.1182/0.1512 | 19.364 | 2655 | $2.2 \times 10^{-16}$ * |
| *Pica pica/Pica hudsonia* | 0.152/0.105 | 0.002 * | 0.1303/0.1054 | 20.181 | 2198 | $2.2 \times 10^{-16}$ * |

Table 1.3: $F_{ST}$ estimated with biallelic one-SNP-per-locus dataset and number of fixed SNPs from both full and thinned dataset. *P*-values are from a *G*-test run with 99 simulations. (Negative $F_{ST}$ is the result of computational idiosyncrasies and is effectively 0.0.)

| | $F_{ST}$ | P-value | Number fixed loci (full dataset) | Number fixed loci (1-SNP-per-locus dataset) |
|---|---|---|---|---|
| **Anseriformes** | | | | |
| *Clangula hyemalis* | 0.0039 | 0.05 | 0 | 0 |
| *Anas crecca* | 0.0191 | 0.01 | 0 | 0 |
| *Anas penelope /A. americana* | 0.0439 | 0.01 | 0 | 0 |
| **Charadriiformes** | | | | |
| *Pluvialis squatarola* | -0.0013 | 0.72 | 0 | 0 |
| *Numenius phaeopus* | 0.269 | 0.01 | 31 | 12 |
| *Tringa brevipes /T. incana* | 0.585 | 0.01 | 299 | 121 |
| **Passeriformes** | | | | |
| *Luscinia svecica* | 0.0138 | 0.03 | 0 | 0 |
| *Pinicola enucleator* | 0.442 | 0.01 | 283 | 91 |
| *Pica pica/Pica hudsonia* | 0.328 | 0.01 | 84 | 35 |

Table 1.4: Model comparisons for each lineage, with two-population models presented first and the single one-population lineage (*Pluvialis squatarola*) below. Maximum log composite likelihood (MLCL) is averaged from the top five runs of the best optimized model presented for each. “n.a.” indicates that the model was unstable and could not be run to completion.

| | **Neutral** | **IM** | **Island** | **Split-migration** | |
|---|---|---|---|---|---|
| **Anseriformes** | | | | | |
| *Clangula hyemalis* | -1238.01 | n.a. | -2841.08 | -245.87 | |
| *Anas crecca* | -1291.50 | n.a. | -1593.81 | -245.40 | |
| *Anas penelope /A. americana* | -1343.25 | -929.99 | -2157.45 | -281.05 | |
| **Charadriiformes** | | | | | |
| *Numenius phaeopus* | -485.17 | -130.42 | -207.77 | -99.12 | |
| *Tringa brevipes /T. incana* | -5643.88 | n.a. | -496.01 | -381.30 | |
| **Passeriformes** | | | | | |
| *Luscinia svecica* | -1074.99 | n.a. | -1936.60 | -200.08 | |
| *Pinicola enucleator* | -2473.34 | n.a. | -151.61 | -126.56 | |
| *Pica pica/Pica hudsonia* | -5036.21 | n.a. | -256.83 | -167.41 | |
| | **Neutral** | **Two-epoch** | **Growth** | **Bottle-growth** | **Three-epoch** |
| ***Pluvialis squatarola*** | -509.87 | -182.18 | -74.59 | -33.13 | -34.42 |

Table 1.5: Results of two-population δaδi analyses, including parameters and interpreted values—presented as interpreted value (parameter)—including effective population sizes ($N_e$) of Alaskan and Russian populations, effective population size of the ancestral population ($N_{ref}$), time since divergence in years, and migration rates using population sizes of Russian ($v_A$) and Alaskan ($v_{NA}$) populations. All estimates include 95% CI calculated with 100 δaδi bootstraps.

| | **$N_e$ Asian** | **$N_e$ North American** | **$N_{ref}$** | **Time since split (years)** | **Migration (from $v_A$)** | **Migration (from $v_{NA}$)** |
|---|---:|---:|---:|---:|---:|---:|
| **Anseriformes** | | | | | | |
| *Clangula hyemalis* | 65,986 (±5,254)<br>(5.441 ± 0.43) | 75,724 (±6,623)<br>(6.244 ± 0.55) | 12,127 (±468)<br>(113.14 ± 4.27) | 247,657 (± 13,401)<br>(2.042 ± 0.11) | 1.26 (±0.065)<br>(0.462 ± 0.02) | 1.44 (±0.073) |
| *Anas crecca* | 140,029 (± 13,256)<br>(6.045 ± 0.57) | 116,665(± 12,639)<br>(5.036 ± 0.54) | 23,165 (± 351)<br>(118.01 ± 1.79) | 188,543(± 4,820)<br>(1.628 ± 0.04) | 0.840 (± 0.037)<br>(0.278 ± 0.01) | 0.700 (± 0.031) |
| *Anas penelope /A. americana* | 123,011 (± 11,851)<br>(4.931 ± 0.47) | 115,693 (± 7,379)<br>(4.638 ± 0.27) | 24,947 (± 563)<br>(116.13 ± 2.62) | 196,399 (± 6,561)<br>(1.437 ± 0.05) | 0.727 (±0.032)<br>(0.295 ± 0.01) | 0.684 (± 0.030) |
| **Charadriiformes** | | | | | | |
| *Numenius phaeopus* | 22,661 (±1,449)<br>(3.244 ± 0.21) | 12,368 (±721)<br>(1.770 ± 0.10) | 6,986 (±208)<br>(177.28 ± 5.27) | 234,124 (± 11,555)<br>(1.511 ± 0.07) | 0.076 (±0.005)<br>(0.0471 ± 0.003) | 0.042 (± 0.0027) |
| *Tringa brevipes /T. incana* | 19,867 (±2,060)<br>(4.507 ± 0.47) | 8,434 (±620)<br>(1.913 ± 0.14) | 4,408 (±525)<br>(79.49 ± 9.46) | 256,430 (± 23,127)<br>(5.965 ± 0.54) | 0.0215 (±0.0014)<br>(0.0095 ± 0.0006) | 0.009 (± 0.0006) |
| **Passeriformes** | | | | | | |
| *Luscinia svecica* | 207,593 (± 16,441)<br>(4.610 ± 0.36) | 138,918 (± 8,559)<br>(3.085 ± 0.19) | 45,027 (± 1,826)<br>(158.56 ± 6.43) | 284,065 (± 13,312)<br>(1.577 ± 0.07) | 1.323 (± 0.076)<br>(0.574 ± 0.033) | 0.885 (± 0.05) |

Table 1.5 continued

| | | | | | | |
|---|---|---|---|---|---|---|
| *Pinicola enucleator* | 32,803 (± 1,854) (1.246 ± 0.07) | 42,485 (± 1,754) (1.614 ± 0.08) | 26,320 (± 658) (270.56 ± 6.77) | 153,995 (± 6,932) (1.463 ± 0.07) | 0.0028 (± 0.0006) (0.0045 ± 0.0010) | 0.004 (± 0.0008) |
| *Pica pica/Pica hudsonia* | 128,078 (±9,928) (3.410 ± 0.26) | 54,899 (±2,998) (1.462 ± 0.08) | 37,561 (±1,458) (162.39 ± 6.31) | 364,903 (± 22,082) (1.943 ± 0.12) | 0.0140 (±0.0022) (0.0082 ± 0.0013) | 0.006 (± 0.001) |

Table 1.6: Results of one-population δaδi analyses, including parameters and interpreted values—presented as interpreted value (parameter)—including effective population sizes before ($N_{ref}$), during, and after a population contraction and the estimated start and end dates of that population contraction. All estimates include 95% CI calculated with 100 δaδi bootstraps.

| | $N_{ref}$ | $N_e$ during population contraction | $N_e$ after population contraction | T population contraction start | T population contraction end |
|---|---|---|---|---|---|
| ***Pluvialis squatarola*** | 73,413 (± 7,187) (266.91 ± 26.13) | 10,077 (± 1,059) (0.1372 ± 0.0144) | 225,682 (± 15,158) (3.074 ± 0.206) | 1,855,511 (±287,007) (1.552 ± 0.240) | 1,858,543 (± 160,760) (1.555 ± 0.134) |

### 1.8 References

Afgan E, Baker D, van den Beek M, *et al.* (2016) The Galaxy platform for accessible, reproducible, and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, **44**, W3-W10.

Arnold TW, Clark RG (1996) Survival and philopatry of female dabbling ducks in southcentral Saskatchewan. *Journal of Wildlife Management*, **3**, 560-568.

Avise J (2004) *Molecular Markers, Natural History, and Evolution*. 2nd ed. Chapman and Hall, New York.

Bejerano G, Pheasant M, Makunin I, *et al.* (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321-1325.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, **30**, 2114-2120.

Cahill JA, Stirling I, Kistler L, *et al.* (2015) Genomic evidence of geographically widespread of gene flow from polar bears into brown bears. *Molecular Ecology,* **24**, 1205-1217.

Casillas S, Barbadilla A (2017) Molecular population genetics. *Genetics*, **205**, 1003-1035.

Claramunt S, Cracraft J (2015) A new time tree reveals Earth history's imprint on the evolution of modern birds. *Science Advances*, **1**, e1501005.

Coyne JA, Orr HA (2004) *Speciation*. Sinauer Associates, Inc, Sunderland, Massachusetts.

Danecek P, Auton A, Abecasis G, *et al*. (2011) The Variant Call Format and VCFtools. *Bioinformatic*s, **27**, 2156-2158.

Delmore KE, Hübner S, Kane NC, *et al.* (2015) Genomic analysis of a migratory divide reveals candidate genes for migration and implicates selective sweeps in generating islands of differentiation. *Molecular Ecology,* **24**, 1873-1888.

Ellegren H, Smeds L, Burri R, *et al.* (2012) The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature*, **491**, 756-760.

Faircloth BC (2013) illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. http://dx.doi.org/10.6079/J9ILL.

Faircloth BC (2016) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics,* **32**,786-788.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology,* **61**, 717–726.

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.

Feder JL, Egan SP, Nosil P (2012) The genomics of speciation-with-gene-flow. *Trends in Genetics*, **28**, 342-350.

Flaxman SM, Wacholder AC, Feder JL, Nosil P (2014) Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23**, 4074-4088.

Funk DJ, Omland KE (2003) Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology, Evolution, and Systematics*, **34**, 397-423.

Gill RE, McCaffery BJ, Tomkovich PS (2002). Wandering Tattler (*Tringa incana*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Glenn TC, Nilsen R, Kieran TJ, *et al.* (2016) Adapterama I: Universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). Accepted at Molecular Ecology Resources, pending minor revisions, available at http://biorxiv.org/content/early/2016/06/15/049114

Good JM, Vanderpool D, Keeble S, Bi K (2015) Negligible nuclear introgression despite complete mitochondrial capture between two species of chipmunks. *Evolution,* **69**, 1961-1972.

Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, **29**, 644-652.

Grant MC (1991) Nesting densities, productivity and survival of breeding Whimbrel Numenius phaeopus in Shetland. *Bird Study*, **38**, 160-169.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, **5**, e1000695.

Guzy MJ, McCaffery (2002). Bluethroat (*Luscinia svecica*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Harvey MG, Smith BT, Glenn TC, Faircloth BC, Brumfield RT (2016) Sequence capture versus restriction site associated DNA sequencing for shallow systematics. *Systematic Biology*, **65**, 910-924.

Hendry AP, Bolnick DI, Berner D, Peichel CL (2009) Along the speciation continuum in sticklebacks. *Journal of Fish Biology*, **75**, 2000-2036.

Hopkins DM (1959) Cenozoic history of the Bering Land Bridge. *Science*, **129**, 1519-1528.

Hopkins DM, ed. (1967) *The Cenozoic history of Beringia—A synthesis*. Stanford University Press, The Bering Land Bridge.

Hopkins DM, MacNeil FS, Merklin RL, Petrov OM (1965) Quaternary correlations across Bering Strait. *Science*, **147**, 1107-1114.

Humphries E, Winker K (2011) Discord reigns among nuclear, mitochondrial, and phenotypic estimates of divergence in nine lineages of Beringian birds. *Molecular Ecology,* **20**, 573-583.

Johnson K (1995) Green-winged Teal (*Anas crecca*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Jombart T, Ahmed I (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, **27**, 3070-3071.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]

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, *et al*. (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics,* **25**, 2078-2079.

Lischer HEL, Excoffier L (2012) PGDSpider: an automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, **28**, 298-299.

McKenna A, Hanna M, Banks E, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* **20**, 1297-1303.

Melles M, Brigham-Grette J, Minyuk PS *et al.* (2012) 2.8 million years of Arctic climate change from Lake El'gygytgyn, NE Russia. *Science*, **337**, 315-320.

Mini AE, Harrington ER, Rucker E, Dugger BD, Mowbray TB (2014) American Wigeon (*Anas americana*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Nosil P, Funk DJ, Ortiz-Barrientos D (2009) Divergent selection and heterogeneous genomic divergence. *Molecular Ecology*, **18**, 375-402.

Nosil P, Feder JL, Flaxman SM, Gompert Z (2017) Tipping points in the dynamics of speciation. *Nature Ecology and Evolution*, **1**, 0001.

Peters JL, McCracken KG, Pruett CL, *et al*. (2012) A parapatric propensity for breeding precludes the completion of speciation in common teal (*Anas crecca* sensu lato). *Molecular Ecology,* **21**, 4563-4577.

Peters J, Winker K, Millam KC *et al*. (2014) Mito-nuclear discord in six congeneric lineages of Holoarctic ducks (genus *Anas*). *Molecular Ecology,* **23**, 2961-2974.

Pielou EC (1991) *After the Ice Age: the return of life to glaciated North America*. University of Chicago Press, Chicago.

Poole AF, Pyle P, Patten MA, Paulson DR (2016) Black-bellied Plover (*Pluvialis squatarola*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Price T (2008) *Speciation in Birds.* Roberts and Company Publishers, Greenwood Village, Colorado.

R Core Team (2015) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ragland GJ, Doellman MM, Meyers P *et al.* (2017) A test of genomic modularity among life history adaptations promoting speciation-with-gene-flow. *Molecular Ecology*, **26**, 3926-3942.

Riesch R, Muschick M, Lindtke D, *et al.* (2017) Transitions between phases of genomic differentiation during stick-insect speciation. *Nature Ecology and Evolution*, **1**, 0082.

Robertson GJ, Savard J-PL (2002) Long-tailed Duck (*Clangula hyemalis*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Roux C, Fraïsse C, Romiguier J, Anciaux Y, Galtier N, Bierne N (2016) Shedding light on the grey zone of speciation along a continuum of genomic divergence. *PLOS Biology*, **14**, e2000234.

Saether BE, Lande R, Engen S, *et al.* (2005) Generation time and temporal scaling of bird population dynamics. *Nature,* **436**, 99-102.

Seehausen O, Butlin RK, Keller I, *et al.* (2014) Genomics and the origin of species. *Nature Reviews Genetics*, **15**, 176-191.

Siepel A, Bejerano G, Pedersen JS, *et al*. (2005) Evolutionary conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, **16**, 164-172.

Skeel MA, Mallory EP (1996) Whimbrel (*Numenius phaeopus*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Sousa V, Hey J (2013) Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, **14**, 404-414.

Stephen S, Pheasant M, Makunin IV, Mattick J (2008) Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Molecular Biology and Evolution*, **25,** 401-408.

Toews DPL, Brelsford A (2012) The biogeography of mitochondrial and nuclear discordance in animals. *Molecular Ecology*, **21**, 3907-3930.

Toews DPL, Taylor SA, Vallender R, Brelsford A, Butcher BG, Messer PW, Lovette IJ (2016) Plumage genes and little else distinguish the genomes of hybridizing warblers. *Current Biology,* **26**, 1-6.

Trost CH (1999) Black-billed Magpie (*Pica hudsonia*). The Birds of North America (P.G. Rodewald, Ed.). Ithaca: Cornell Lab of Ornithology. Retrieved from the Birds of North America Online.

Van Belleghem SM, Rastas P, Papanicolaou A, *et al.* (2017) Complex modular architecture around a simple toolkit of wing pattern genes. *Nature Ecology and Evolution*, **1**, 0052.

Via S, West J (2008) The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, **17**, 4334-4345.

Vijay N, Weissensteiner M, Burri R, Kawakami T, Ellegren H, Wolf JBW (2017) Genome-wide patterns of variation in genetic diversity are shared among populations, species and higher order taxa. *Molecular* Ecology, **26**, 4284-4295.

Winker K, McCracken KG, Gibson DD, Peters JL (2013) Heteropatric speciation in a duck, *Anas crecca*. *Molecular Ecology*, **22**, 5922-5935.

Wolf JBW, Ellegren H (2017) Making sense of genomic islands of differentiation in light of speciation. *Nature Reviews Genetics*, **18**, 87-100.

Zarza E, Faircloth BC, Tsai WLE, Bryson Jr RW, Klicka J, McCormack JE (2016) Hidden histories of gene flow in highland birds revealed with genomic markers. *Molecular Ecology*, **25**, 1-14.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203-214.

## 1.9 Appendix to Chapter 1: Supplemental Tables

Table S1. Specimen numbers for samples used in this study and associated data archiving information. UAM = University of Alaska Museum, UWBM = University of Washington Burke Museum.

| | Asian Samples | North American Samples | Localities | Sequence Read Archive |
|---|---|---|---|---|
| ***Clangula hyemalis*** | UWBM 43893, UWBM 43894, UWBM 43895, UWBM 43916, UWBM 43917, UWBM 43918, UWBM 43919, UWBM 43970 | UAM 13154, UAM 21883, UAM 21931, UAM 25746, UAM 28184, UAM 29027, UAM 29029, REW 564 | Anadyr’ (8); Dalton Hwy (7), Barrow (1) | |
| ***Anas crecca*** | UAM 9255, UAM 9191, UAM 11334, UAM 11335, UAM 14100, UAM 14666, UAM 22853 | UAM 11251, UAM 11920, UAM 14961, UAM 20635, UAM 24497, UAM 28085, UAM 28444 | Shemya I. (6), Yakutia (1) ; Izembek NWR (3), Chirikof I. (1), Eielson AFB (1), Fairbanks (1), Monashka Bay (1) | |
| ***Anas penelope /A. americana*** | UAM 8758, UAM 9359, UAM 10008, UAM 11803, UAM 24301, UAM 24455, UAM 24550, UAM 27749 | UAM 11908, UAM 11909, UAM 11916, UAM 11919, UAM 13141, UAM 26061, UAM 28087, UAM28088, | Shemya I. (6), Attu I. (1), Buldir I. (1) ; Fairbanks (5), Deadman Bay (1), Eielson AFB (1), Monashka Bay (1) | |
| ***Pluvialis squatarola*** | UWBM 43931, UWBM 43963, UWBM 43964, UWBM 44550, UWBM 51608 | UAM 14237, UAM 14238, UAM 14239, UAM 14240, UAM 14241 | Anadyr’ (3), Zaliv Odyan (1), Cherskiy (1); Mactan I. (4), Olango I. (1). | |
| ***Numenius phaeopus*** | UAM 8212, UAM 14225, UAM 14229, UAM 14230, UAM 14625, UAM 21379, UAM 28213, UAM 28214 | UAM 9260, UAM 11507, UAM 13349, UAM 13925, UAM 14928, UAM 17980, UAM 20642, UAM 28602 | Olango I. (3), Shemya I. (2), Adak I. (1), Attu I. (1), Buldir I. (1); Seward Pen. (4), Dalton Hwy (1), Goodnews Bay (1),Ugashik Bay (1), Ugak I. (1). | |

Table S1 continued

| | | | |
|---|---|---|---|
| ***Tringa brevipes /T. incana*** | UAM 7534, UAM 8521, UAM 8805, UAM 10112, UAM 28427, UAM 28428, UAM 28429, UAM 28430 | UAM 8240, UAM 13434, UAM 13569, UAM 21813, UAM 24859, UAM 28422, UAM 28425, UAM 28426 | Shemya I. (5), Attu I. (3); Attu I. (5), Amlia I. (1), Ban I. (1), Kodiak I. (1) |
| ***Luscinia svecica*** | UWBM 44233, UWBM 44242, UWBM 44246, UWBM 44361, UWBM 44362, UWBM 44363, UWBM 44629, UWBM44630 | UAM 8584, UAM 8585, UAM 8944, UAM 8945, UAM 8946, UAM 15419, UAM 17727 | Milkovo (4), Cheriskiy (3); Taylor Hwy. (5), Dalton Hwy. (2) |
| ***Pinicola enucleator*** | UAM 24601, UAM 24602, UWBM 44628, UWBM 47314, UWBM 47315, UWBM 51627, UWBM 51642, UWBM 51643 | UAM 8794, UAM 8797, UAM 10158, UAM 11286, UAM 11287, UAM 26361, UAM 28530 | Snezhnaya Dolina (3), Sakhalinskaya Oblast’ (2), Shemya I. (2), Kamchatka (1); Kodiak I. (5), Revillagigedo I. (2) |
| ***Pica pica/Pica hudsonia*** | UWBM 44581, UWBM 44584, UWBM 44585, UWBM 47197, UWBM 72084, UWBM 72091, UWBM 74569 | UAM 8511, UAM 10105, UAM 12453, UAM 13052, UAM 13053, UAM 17742, UAM 27024 | Kamchatka (3), Ussuriysk (2), Gayvoron (1), Khabarovskiy Kray (1); Izembek NWR (3), Chirikof I. (1), Cold Bay (1), Kodiak I. (1), Popof I. (1) |

Table S2. Genome assembly used for calculation of substitution rates and time since most recent common ancestor (TMRCA, from Claramunt & Cracraft 2015), substitution rate, and generation time for each lineage.

| | **Genome** | **Genbank accession number** | **TMRCA (my)** | **Substitution rate (subs/site/gen)** | **Generation time (yrs)** | **Sources** |
|---|---|---|---|---|---|---|
| **Anseriformes** | | | | | | |
| *Clangula hyemalis* | *Anser cygnoides* | GCA_000971095.1 | 28.1329 | 2.270 x $10^{-9}$ | 5 | Robertson & Savard 2002 |
| *Anas crecca* | *Anser cygnoides* | GCA_000971095.1 | 28.1329 | 1.224 x $10^{-9}$ | 2.5 | Arnold & Clark 1996; Johnson 1995. |
| *Anas penelope /A. americana* | *Anser cygnoides* | GCA_000971095.1 | 28.1329 | 1.084 x $10^{-9}$ | 2.74 | Arnold & Clark 1996; Mini *et al.* 2014. |
| **Charadriiformes** | | | | | | |
| *Pluvialis squatarola* | *Calidris pugnax* | GCA_001458055.1 | 53.5153 | 2.426 x $10^{-9}$ | 8.14 | Poole *et al.* 2016 |
| *Numenius phaeopus* | *Charadrius vociferus* | GCA_000708025.2 | 53.5153 | 4.418 x $10^{-9}$ | 11.09 | Grant 1991; Skeel & Mallory 1996 |
| *Tringa brevipes /T. incana* | *Charadrius vociferus* | GCA_000708025.2 | 53.5153 | 2.453 x $10^{-9}$ | 4.88 | Gill *et al.* 2002 |
| **Passeriformes** | | | | | | |
| *Luscinia svecica* | *Sturnus vulgaris* | GCA_001447265.1 | 30.9672 | 9.6896 x $10^{-10}$ | 2 | Guzy & McCaffery 2002; |
| *Pinicola enucleator* | *Zonotrichia albicollis* | GCA_000385455.1 | 21.7098 | 1.577 x $10^{-9}$ | 2 | |
| *Pica pica/Pica hudsonia* | *Taeniopygia guttata* | GCA_000151805.2 | 41.5132 | 1.304 x $10^{-9}$ | 2.5 | Trost 1999 |

Table S3. Overall results of assembly of four reference individuals in each lineage using Trinity v 2.0.6, including number of contigs, total bp, average length, minimum and maximum length, and number of contigs over 1 kb.

| | Contigs | Total bp | Average length (bp) | Min-max length | Contigs > 1 Kb |
|---|---|---|---|---|---|
| **Anseriformes** | | | | | |
| *Clangula hyemalis* | 361,602 | 124,386,185 | 344 | 201-10,795 | 4,037 |
| *Anas crecca* | 446,548 | 163,059,294 | 365 | 224-16,935 | 4,490 |
| *Anas penelope /A. americana* | 130,506 | 47,215,417 | 362 | 224-16,647 | 1,653 |
| **Charadriiformes** | | | | | |
| *Pluvialis squatarola* | 261,906 | 82,685,953 | 316 | 201-16,980 | 1,086 |
| *Numenius phaeopus* | 443,111 | 153,588,413 | 347 | 201-15,750 | 3,823 |
| *Tringa brevipes /T. incana* | 244,794 | 82,185,517 | 336 | 201-18,515 | 2,925 |
| **Passeriformes** | | | | | |
| *Luscinia svecica* | 460,117 | 159,588,295 | 347 | 201-16,815 | 2,985 |
| *Pinicola enucleator* | 657,330 | 254,336,867 | 387 | 201-17,196 | 9,935 |
| *Pica pica/Pica hudsonia* | 386,147 | 130,926,174 | 339 | 201-17,217 | 2,634 |

Table S4. UCE reference assemblies used in each lineage, including number of UCEs, total bp, average length, minimum and maximum length, and number of contigs over 1 kb.

| | **UCEs identified** | **Total bp** | **Average length (bp)** | **Min - max length** | **Contigs > 1 Kb** |
|---|---|---|---|---|---|
| **Anseriformes** | | | | | |
| *Clangula hyemalis* | 4,154 | 4,117,721 | 991 | 202-2,069 | 2,037 |
| *Anas crecca* | 4,076 | 4,024,944 | 987 | 231-11,107 | 1,945 |
| *Anas penelope /A. americana* | 4,154 | 3,410,066 | 821 | 225-1,740 | 648 |
| **Charadriiformes** | | | | | |
| *Pluvialis squatarola* | 4,294 | 2,890,550 | 673 | 226-1,842 | 63 |
| *Numenius phaeopus* | 4,257 | 4,108,788 | 965 | 202-2,237 | 1,897 |
| *Tringa brevipes /T. incana* | 4,251 | 3,830,221 | 901 | 204-2,435 | 1,399 |
| **Passeriformes** | | | | | |
| *Luscinia svecica* | 4,040 | 3,466,742 | 858 | 201-2,094 | 870 |
| *Pinicola enucleator* | 4,244 | 5,226,843 | 1,232 | 217-2,771 | 3,249 |
| *Pica pica/Pica hudsonia* | 4,286 | 3,749,321 | 875 | 216-1,945 | 940 |

## Chapter 2: An empirical examination of the effects of sample size variation on population demographic estimates in nonmodel organisms using single nucleotide polymorphism data[1]

### 2.1 Abstract

Sample size is a critical aspect of study design in population genomics research, yet few empirical studies have examined the impacts of small sample sizes. We used eight datasets of ultraconserved elements (UCEs) making pairwise comparisons of bird populations showing different levels of divergence (populations, subspecies, and species). All individuals were genotyped at all loci. We estimated population demographic parameters of effective population size, migration rate, and time since divergence using Diffusion Approximation for Demographic Inference (δaδi), an allele frequency spectrum (AFS) method. Using split-migration models optimized with full datasets, we subsampled at sequentially smaller sample sizes from full datasets of 6 – 8 diploid individuals per population and then compared estimates and their variances. Effective population size parameters (*v*) tended to be underestimated at low sample sizes (fewer than 3 diploid individuals per population), migration (*m*) was fairly reliably estimated until 2 individuals per population, and no trend of over- or underestimation was found in either time since divergence (*T*) or $\Theta$ ($4N_{ref}\mu$). Lineages that were split above the population level (subspecies and species pairs) tended to have lower variance at smaller sample sizes than population-level splits, with many parameters reliably estimated down to 3 individuals per population, whereas population-level splits often required at least 5 individuals per population for reliable demographic inferences. Although divergence levels may be unknown at the outset of study design, our

[1] McLaughlin JF, Faircloth BC, Glenn TC, Winker K. An empirical examination of the effects of sample size variation on population demographic estimates in nonmodel organisms using single nucleotide polymorphism data. Prepared for *Molecular Ecology Resources.*

results provide a framework for planning appropriate sampling, and for interpreting results if smaller sample sizes must be used.

## 2.2 Introduction

Next-generation sequencing (NGS) has created a massive increase in the quantity of data available for studying population histories. Increased numbers of loci improve the resolution of demographic estimates (Jeffries *et al.* 2016, Nazareno *et al.* 2017), including effective population size, migration rate, and time since divergence, even when the number of sampled individuals in a given population is relatively low (Willing *et al.* 2012, Jeffries *et al.* 2016, Nazareno *et al.* 2017). However, it is not well understood how the precision and accuracy of these estimates are impacted by relatively low population sample sizes. The number of individuals per population able to be included in a study may still be limited by factors such as availability of samples for isolated, difficult-to-access populations (Pruett & Winker 2008), tradeoffs between including more individuals per population or more populations, and decisions of whether to include more loci or more individuals (Felsenstein 2005, Jeffries *et al.* 2016). Because these issues affect study design, it is important to understand the impacts of relatively low population sample sizes on commonly estimated population demographic parameters.

The impacts of population sample size, and particularly the tradeoff between increased numbers of individuals versus increased numbers of loci, has been studied previously, primarily with microsatellite datasets. In general, increasing the number of loci decreases the number of individuals required for accurate parameter estimations in population genetic studies (Morin *et al.* 2009, Willing *et al*. 2012), but different parameter estimates are affected differently at low sample sizes. A size of 8 alleles per population (4:4 diploid individuals), has been suggested as an optimum sample size for obtaining coalescent-based likelihood estimates of $\Theta = 4N_e\mu$ (Felsenstein 2005). This sample size has also been sufficient for non-coalescent-based estimates of unbiased heterozygosity (Pruett & Winker 2008), which have been effectively estimated with 5 – 10 individuals. However, other estimators such as genetic diversity and differentiation ($F_{ST}$) require larger sample sizes, and often the number of individuals required for accurate estimates increases as divergence decreases (Kalinowski 2005, Morin *et al.* 2009)

NGS datasets, with their large increases in numbers of loci sampled, are predicted to decrease the number of individuals required for obtaining accurate estimates of demographic history (Jeffries *et al.* 2016). However, impacts of sample size on such estimates have undergone only limited investigation thus far, and previous empirical work has focused on estimates of diversity ($A_E$, $H_O$, and unbiased $H_E$) and differentiation ($F_{ST}$; Nazareno *et al.* 2017). Other demographic estimates made using allele frequency spectrum (AFS) methods have only been evaluated so far with simulation data (Robinson *et al.* 2014), using the program δaδi (Diffusion Approximation for Demographic Inference; Gutenkunst *et al.* 2009). These authors showed that median estimated parameter values in two-population δaδi models of divergence in isolation remained close to true values down to 3 diploid individuals per population, but this did not hold true across all three model types that they examined, and their optimal sampling recommendations depended on the timescale of the demographic events experienced by the populations, with very recent and very ancient events both requiring greater sample sizes (Robinson *et al*. 2014). In empirical systems, such information on the timescale of demographic events or divergence may be unknown at the outset of a study, particularly in taxa that have not been studied, and care must be taken to avoid sampling too few individuals to accurately estimate parameters of interest.

Here we use empirical datasets to examine how inferences of population parameters are impacted by sample size, scaling downward from full datasets that meet or exceed sample sizes widely considered optimal for coalescent-based analyses. We expected that as sample sizes decrease, variability of estimates would increase, but to varying degrees among different parameters, and that estimates of parameters might be over- or underestimated at low sample sizes. We use datasets that reflect multiple demographic and evolutionary histories, and explore empirical factors that could impact the ability to accurately infer demographic histories at low sample sizes.

## 2.3 Methods

### *2.3.1 Study system*

We used 8 empirical datasets of ultraconserved elements (UCEs) from Beringian birds, as described in Chapter 1, to generate repeated subsampled datasets at decreasing sample sizes for analysis in δaδi. These datasets represent population, subspecies, and species pairs in three avian orders: Asian and North American populations of *Clangula hyemalis* (long-tailed duck), *Anas crecca crecca/Anas crecca carolinensis* (green-winged teal), and *Anas penelope/Anas americana* (Eurasian and American wigeons) in Anseriformes; *Pluvialis squatarola* (black-bellied plover), *Numenius phaeopus variegatus/Numenius phaeopus hudsonicus* (whimbrel), and *Tringa brevipes/Tringa incana* (gray-tailed and wandering tattlers) in Charadriiformes; and *Luscinia svecica* (bluethroat), *Pinicola enucleator kamschatkensis/Pinicola enucleator flammula* (pine grosbeak), and *Pica pica/Pica hudsonia* (Eurasian and black-billed magpies) in Passeriformes. These datasets, which span divergence levels from populations with a high level of gene flow to almost completely reproductively isolated groups, will enable us to explore how the effects of low sample sizes on demographic inference play out across the speciation continuum.

### *2.3.2 Laboratory procedures*

We extracted whole genomic DNA from 6-8 individuals each from North American and Asian populations of the above taxa, using archived museum specimens (Table S1 in Chapter 1). In brief, we then produced double-indexed libraries of 5,060 UCE probes to subsample the genome, following Glenn *et al.* (2016), which were sequenced using an Illumina HiSeq 2500. More details are given in Chapter 1.

### *2.3.3 Bioinformatics*

Sequencing data were demultiplexed with bcl2fastq (v 1.8.4; Illumina, Inc.). We used illumiprocessor (Faircloth 2013), a parallel wrapper around Trimmomatic (Bolger *et al.* 2014), to trim adapters and low-quality bases. The singleton and read 1 fastq files for each individual were combined and then, with read 2 files, were assembled with Trinity (v2.0.6; Grabherr *et al.* 2011) on Galaxy (Afgan *et al.* 2016). We extracted UCE loci from these assemblies and a complete matrix was constructed for each lineage using PHYLUCE (v 1.5; Faircloth 2016) to determine the median number of loci shared and unshared by individuals in a given dataset.

To build a reference to call SNPs against, in each lineage the fastq sequence files for the two individuals in each population with closest to the median number of unshared loci were combined to produce a single read 1 and read 2 file for each and assembled with Trinity on Galaxy, from which a fasta file of UCE loci for each lineage was created with PHYLUCE. We then indexed these with BWA and SAMtools (Li & Durbin 2009; Li *et al.* 2009) for SNP calling.

We used a modified workflow for population genomics with UCEs developed by Faircloth and Michael Harvey (https://github.com/mgharvey/seqcap_pop) to call SNPs. For each lineage, BWA-MEM (Li 2013) was used to align sequences to the reference. The resulting SAM alignments were converted to BAM with SAMtools. We used Picard to check alignments for BAM format violations, add read-groups header information, and mark PCR duplicates for each individual (http://broadinstitute.github.io/picard). All individuals were merged into a single file for each lineage with Picard, then indexed with SAMtools. We then used Genome Analysis Toolkit (GATK; v 3.4-0; McKenna *et al.* 2010) to find and realign around indels, then called SNPs using the UnifiedGenotyper tool within GATK. SNPs and indels were then annotated and indels masked. We restricted our datasets to high-quality SNPs (Q30) and performed read-backed phasing. We then used VCFtools (Danecek *et al.* 2011) to reduce our datasets to a complete matrix for each lineage with a minimum genotype quality (GQ) of 10. We thinned each dataset to one SNP per locus with VCFtools to minimize the impact of linkage (as δaδi requires unlinked SNPs), then removed Z-linked loci with a custom Python script due to their differing inheritance pattern (https://github.com/jfmclaughlin92/thesis).

### *2.3.4 Subsampling datasets and analyses*

To produce datasets of varying sample sizes, stepping down from the maximum number of individuals available for each population (6 – 8) to 1 individual per population, a custom Python script (https://github.com/jfmclaughlin92/thesis) was used. This script iteratively sampled individuals without replacement from the thinned .vcf files, created new .vcf files containing these individuals, converted these files to the proper δaδi input format (using a Perl script by Kun Wang;

https://groups.google.com/forum/#!msg/dadi-user/p1WvTKRI9_0/1yQtcKqamPcJ), and ran δaδi models with pre-determined best-fit parameters for a split-migration model. For each sample size, 25 subsampled datasets were created, which were each run five times. The best fit run by highest maximum log composite likelihood (MLCL) value among those five runs was then selected for each dataset and used for subsequent analyses. Parameter estimates for effective population size ($v_1$ and $v_2$), migration (*m*), time since split (*T*), and $\Theta$, defined as $4N_{ref}\mu$, with $N_{ref}$ defined as ancestral population size and $\mu$ as mutation rate per generation, were then compared across different sample sizes. The root mean square error (SRMSE) was calculated, defined as

$$SRMSE_{\theta} = \sqrt{\frac{\sum(\hat{\theta} - \theta)^2}{n}} \Big/ \bar{\theta} \tag{2.1}$$

with $\theta$ in this context representing the estimate from the full dataset, $\hat{\theta}$ as the parameter estimate from the subsampled dataset, and *n* the number of datasets (25) considered, as following Robinson *et al.* (2014). This was then scaled by the mean of the parameter estimate at each sample size ($\bar{\theta}$) to allow inter-lineage comparisons of the changes in accuracy at lower sample sizes (SRMSE). This enables us to quantify the changes in accuracy of estimates at different sample sizes relative to each species' parameter estimates' means.

## 2.4 Results

Each lineage had a dataset of between 1,636 and 2,656 variable loci (Table 2.1). Across the eight lineages, 25 datasets were constructed at each sample size, starting at one individual per population and increasing to one less than the full sample size for a total of 1,250 subsampled datasets.

Overall, as expected, variance in parameter estimates increased with smaller sample sizes (Table 2.2). Performance of mean parameter estimates varied with lineage and sample size. The effective population size parameters ($v_1$ and $v_2$) were routinely underestimated at lower sample sizes, whereas there was a trend towards overestimation of migration (*m*) (Table 2.2; Figure 2.1). The other two parameters, *T* and *Θ*, were more ambiguous, with both over- and under-estimation occurring in different lineages (Figure 2.1). These corresponded to large changes in the biologically meaningful estimates derived from these parameters. This can be seen in the effective population size parameter of *Tringa brevipes* ($v_1$), which varied from 1.016 to 8.487 across sample sizes (Table 2.2). In biologically meaningful terms, this represented effective population size estimates of 4,478 to 37,410 individuals.

In general, SRMSE increased as sample sizes decreased (Table 2.3), indicating the loss of accuracy at lower sample sizes. Additionally, lineages with less divergence tended to exhibit more variance at higher sample sizes than deeper splits (Figure 2.1), although this was most notable in the two population-level splits (*L. svecica* and *C. hyemalis*; see appendix). In the deeper splits (subspecies/species)—particularly *T. brevipes*/*T. incana*, *N. phaeopus*, and *Pica pica*/*Pica hudsonia*—most parameter estimates reached a similar level of variance at approximately 4 or 5 diploid individuals, at which point adding more individuals did not considerably decrease variance, while SRMSE only began to increase markedly below

3:3 comparisons (Table 2.3). In some shallower splits, such as *A. crecca* and *L. svecica*, SRMSE began increasing in most parameters below a sample size of 5 (Table 2.3). However, this was not universally the case, with SRMSE values in *C. hyemalis* remaining similar at most sample sizes for multiple parameter estimates.

## 2.5 Discussion

Sample size is an important consideration in study design, but it remains understudied in NGS datasets (Nazareno *et al.* 2017). Our results suggest that the minimum reliable sample size will vary by lineage, depending heavily on factors such as parameters of interest and divergence level. Although estimates from coalescent theory have suggested that sample sizes of 8- 10 individuals per population are optimal (Felsenstein 2005), by genotyping both alleles of diploid animals sample sizes were doubled, and we were able to estimate population parameters at considerably lower sample sizes. Certain parameters, such as migration rate (*m*), and effective population sizes ($v_1$, and $v_2$), showed fairly reliable patterns of over- or under-estimation across all lineages (Figure 2.1). In particular, *m* was routinely estimated with relatively low variance down to two individuals per population, below which it was overestimated in all lineages (Table 2; Figure S3). The effective population sizes (*v* parameters) were not as robust, with variance tending to begin to increase markedly below 4 diploid individuals per population and mean estimates decreasing in all lineages. They were, however, still reasonably accurate at relatively small samples sizes (Figures S1, S2). These changes in parameter estimates corresponded with considerable changes in biologically meaningful values, as seen in the change in the estimates of effective population sizes in *T. brevipes* from 4,478 to 37,410 individuals across datasets, demonstrating the impact that small sample sizes can have on estimates.

Our results reinforce previous findings (Kalinowski 2005, Morin *et al.* 2009) that an important factor in determining the minimum sample size for a study is the divergence in the lineages under examination. Although in many cases this may be known at the start of a study, this may not always be true, potentially complicating the determination of sampling design. However, some general recommendations are possible. Lineages with considerable divergence (e.g. species-level, such as in *Tringa*) had accurate demographic parameters estimated at lower sample sizes. Thus it is possible in such systems to reliably use fewer individuals. In population-level splits that may experience substantial gene flow, however, higher sample sizes may be required to overcome the impact of individuals with varying amounts of admixture.

Our findings of the effects of divergence levels on the minimum sample sizes needed to accurately estimate population demographic parameters broadly agreed with previous findings in other genetic markers, with some exceptions. In lineages that are more shallowly split and have experienced more gene flow, greater sample sizes are required to reliably estimate multiple parameters, including not just the demographic parameters examined here, but also in genetic distance (Kalinowski 2005) and $F_{ST}$ (Morin *et al.* 2009, Humphries & Winker 2011). The two population-level splits, *L. svecica* and *C. hyemalis*, did not perform as well in most parameters at sample sizes below 6 individuals per population, with accuracy (as measured by increasing SRMSE) decreasing rapidly, whereas most deeper splits had slower increases in SRMSE as sample size decreased (Table 2.3). The presence of a substantial amount of gene flow appears to increase variance and decrease accuracy, as seen in *L. svecica* (Table 2.2, Table 2.3), and in practical terms would require increased sample sizes for accurate parameter estimation.

Due to computational restrictions, we analyzed all data under the optimal model determined for the full dataset in each lineage and did not investigate the impact of sample size on model fit. Several datasets (notably *Clangula hyemalis*, *A. penelope/A. americana*, and *Luscinia svecica*) showed signs in some parameters of beginning to routinely push the upper bounds of the model. This means that both variance and over-estimation of the parameters was likely underestimated in these groups. This situation has also been noted with simulation data, which have been found in some situations to have a better fit with a model type different than the one under which they were simulated (Robinson *et al.* 2014).

Research efficiency requires attention not only to the minimum sample size required to meet an objective, but also to the point after which adding more samples begins to produce diminishing returns. In this context, this means the point above which the SRMSE becomes similar between sample sizes, but before the means of estimates start to change due to decreased sample size. This inflection point may represent the minimum reliable sample size, but not necessarily. In some lineages, SRMSE was very similar at larger sample sizes, began to slowly increase at intermediate sizes, and then at low sample sizes increased quickly (Figure 2.2). This again varied among lineages (Table 2.3; Appendix 2). In some, such as the *Pica* and *Tringa* species lineages, this inflection point was reached at higher sample sizes than the minimum reliable sample sizes in some parameters, whereas in others, such as in most estimates of *m* (migration), these points were the same (Figure S3). However, in some groups, particularly estimates of $v_1$ (effective population size) and *m* in *L. svecica*, this optimal point was not reached until the full dataset was analyzed, and may not have in fact been reached at all in *C. hyemalis* in any of the parameter estimates (Figures S1-S5). This is consistent with the findings of Robinson *et al.*

(2014), in that although in some cases a small samples size could be used, larger sample sizes still led to more accurate parameter estimates. This was especially the case in these data for *T*, *Θ*, and in some *v* estimates (Table 2.2, Figure 2.1; Figures S1-S5).

Sample size is a critical aspect of study design, and balancing the need for reliable estimates with cost effectiveness is a key tradeoff. Inadequate sampling can lead to ambiguous or biased results (Nazareno & Jump 2012, Nazareno *et al.* 2017), while many parameter estimates are not improved past a given sample size (Felsenstein 2005, Nazareno *et al.* 2017). We found that inference of demographic parameters can be strongly influenced by sample size, with estimates becoming less accurate at lower sample sizes and being over- and underestimated (e.g. Figure 2.2). In general, for pairwise comparisons at the population level, care should be taken to include adequate samples, with the best performance generally occurring with at least 6 or more diploid individuals per population. Parameter estimates in lineages with deeper splits (subspecies and species) were generally more resilient to lower sample sizes; however, this can be confounded by factors such as population structure, gene flow, and related individuals, in which cases sample sizes below 4:4 may not be advisable.

## 2.6 Figures

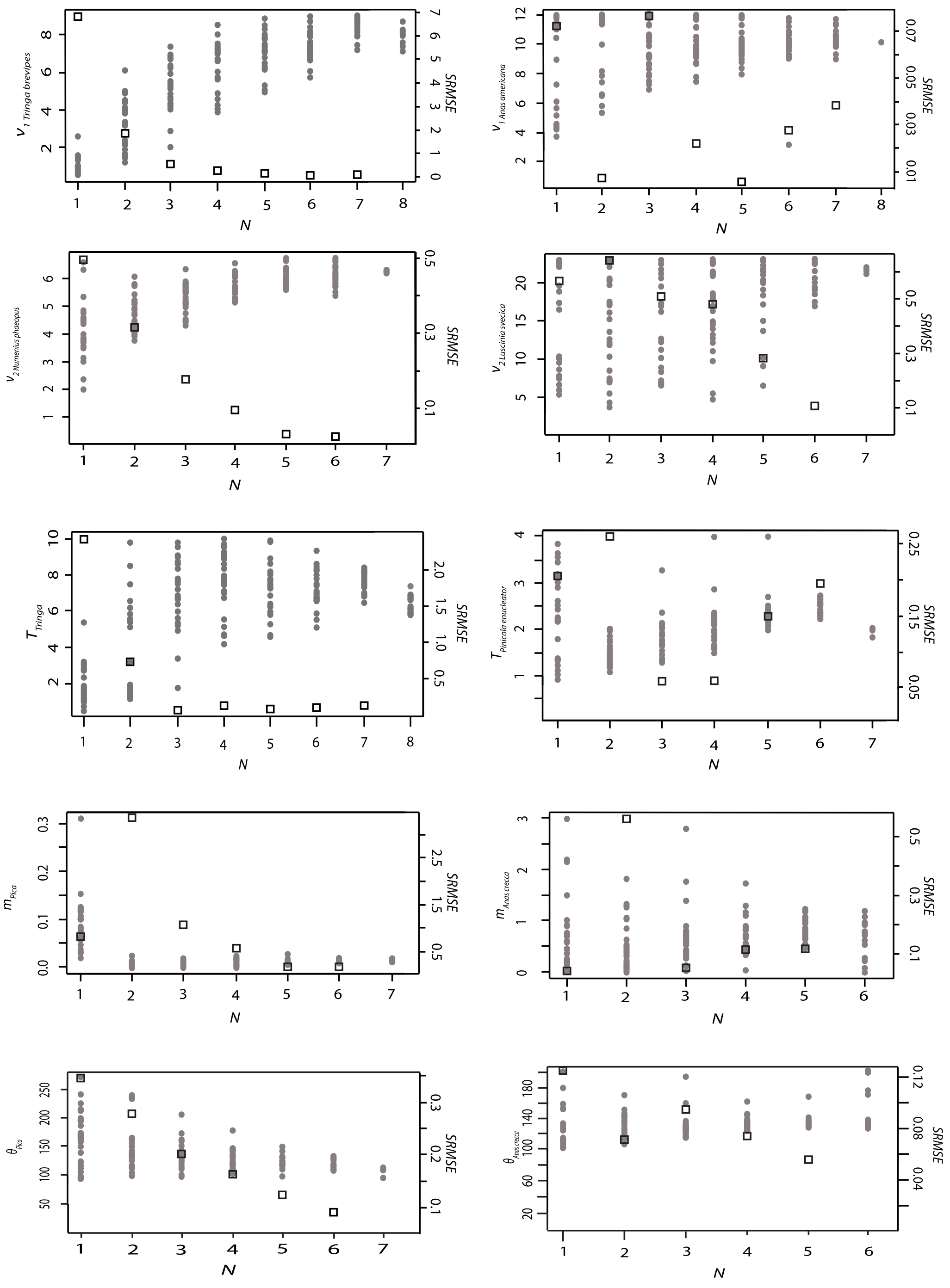

Figure 2.1: Selected results for parameter estimates at varying sample sizes. Parameter estimates of effective population size ($\nu_1$ and $\nu_2$), time since divergence (T), migration (m), and Θ (DEFINE) for selected lineages (in gray; means in black), with scaled root mean square error (SRMSE) on the right axis, indicated with square points.

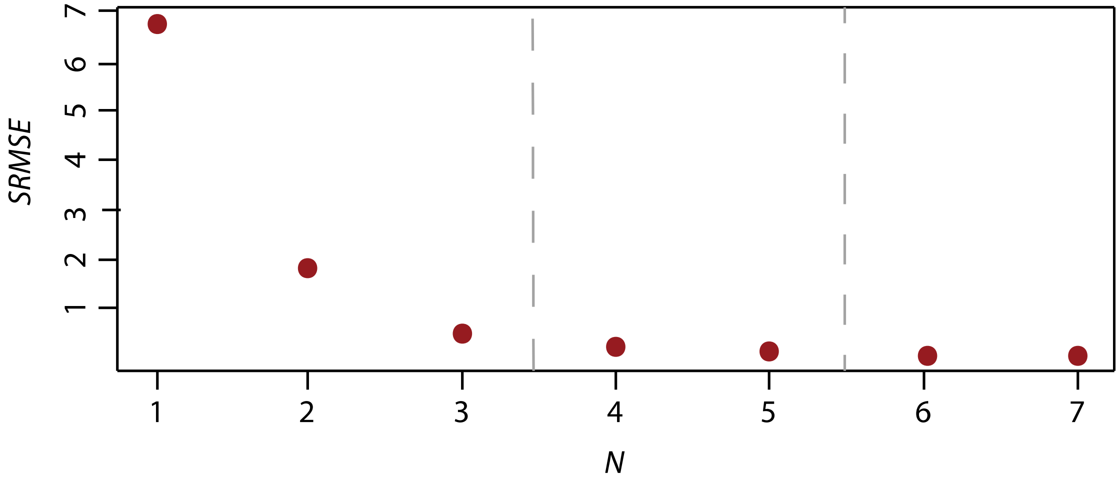

Figure 2.2: SRMSE values for effective population size parameter $v_1$ in T. brevipes. At N = 6 and 7, SRMSE is similar, and sample sizes of these and 8 (the full dataset size) diploid individuals per population yield similar estimates of the parameter without decreasing accuracy markedly. SRMSE increases below this, but sample sizes of 4 and 5 still show only moderately reduced precision. However, below this, SRMSE begins to increase more quickly, and sample sizes of 3 or less have greater accuracy.

### 2.7 Tables

Table 2.1: Number of variable loci in each lineage, and the full dataset size (number of diploid individuals per population).

| | Variable loci | Full dataset size |
|---|---|---|
| **Anseriformes** | | |
| *Clangula hyemalis* | 2,442 | 7 |
| *Anas crecca* | 2,481 | 6 |
| *Anas penelope /A. americana* | 2,315 | 8 |
| **Charadriiformes** | | |
| *Numenius phaeopus* | 2,388 | 7 |
| *Tringa brevipes /T. incana* | 1,636 | 8 |
| **Passeriformes** | | |
| *Luscinia svecica* | 2,516 | 7 |
| *Pinicola enucleator* | 2,656 | 7 |
| *Pica pica/Pica hudsonia* | 2,199 | 7 |

Table 2.2: Mean estimates (± SEM) of demographic parameters $v_1$, $v_2$, *m, T,* and *Θ* in eight lineages of trans-Beringian birds calculated from 25 resampled datasets at each sample size.

| | Parameter | 8:8 | 7:7 | 6:6 | 5:5 | 4:4 | 3:3 | 2:2 | 1:1 |
|---|---|---|---|---|---|---|---|---|---|
| **Anseriformes** | | | | | | | | | |
| *Clangula hyemalis* | $v_1$ | - | 8.937 | 10.706 | 11.039 | 10.662 | 10.977 | 10.688 | 8.864 |
| | | | (± 1.068) | (± 0.449) | (± 0.327) | (± 0.319) | (± 0.234) | (± 0.275) | (± 0.532) |
| | $v_2$ | - | 6.410 | 10.704 | 10.657 | 10.634 | 11.546 | 9.915 | 9.851 |
| | | | (± 1.012) | (± 0.255) | (± 0.318) | (± 0.388) | (± 0.130) | (± 0.344) | (± 0.525) |
| | *T* | - | 1.487 | 1.542 | 1.460 | 1.497 | 1.472 | 1.639 | 2.155 |
| | | | (± 0.213) | (± 0.065) | (± 0.067) | (± 0.083) | (± 0.053) | (± 0.105) | (± 0.187) |
| | *m* | - | 1.217 | 1.524 | 1.554 | 1.704 | 1.847 | 2.093 | 2.324 |
| | | | (± 0.229) | (± 0.121) | (± 0.137) | (± 0.148) | (± 0.143) | (± 0.190) | (± 0.157) |
| | *Θ* | - | 204.806 | 136.062 | 140.407 | 139.646 | 133.497 | 129.653 | 116.837 |
| | | | (± 33.285) | (± 4.133) | (± 6.721) | (± 6.591) | (± 2.999) | (± 4.928) | (± 6.160) |
| *Anas crecca* | $v_1$ | - | - | 13.529 | 13.515 | 13.801 | 12.598 | 13.261 | 11.129 |
| | | | | (± 0.268) | (± 0.229) | (± 0.380) | (± 0.516) | (± 0.526) | (± 0.722) |
| | $v_2$ | - | - | 16.737 | 16.689 | 16.523 | 16.939 | 15.270 | 11.631 |
| | | | | (± 0.450) | (± 0.471) | (± 0.492) | (± 0.516) | (± 1.061) | (± 1.090) |
| | *T* | - | - | 1.154 | 1.226 | 1.265 | 1.333 | 1.298 | 1.500 |
| | | | | (± 0.039) | (± 0.019) | (± 0.024) | (± 0.045) | (± 0.046) | (± 0.088) |
| | *m* | - | - | 0.736 | 0.83 | 0.661 | 0.699 | 0.472 | 0.765 |
| | | | | (± 0.063) | (± 0.040) | (± 0.073) | (± 0.114) | (± 0.094) | (± 0.147) |
| | *Θ* | - | - | 143.00 | 135.50 | 133.17 | 130.67 | 133.51 | 127.13 |
| | | | | (± 4.157) | (± 1.492) | (± 1.581) | (± 3.204) | (± 2.831) | (± 5.231) |

Table 2.2 continued

| | Parameter | 8:8 | 7:7 | 6:6 | 5:5 | 4:4 | 3:3 | 2:2 | 1:1 |
|---|---|---|---|---|---|---|---|---|---|
| *Anas penelope/A. americana* | $v_1$ | 10.116<br>(± 0.002) | 10.518<br>(± 0.132) | 9.847<br>(± 0.318) | 10.063<br>(± 0.217) | 9.904<br>(± 0.260) | 9.398<br>(± 0.320) | 10.193<br>(± 0.466) | 9.438<br>(± 0.618) |
| | $v_2$ | 15.608<br>(± 0.004) | 15.147<br>(± 0.192) | 14.895<br>(± 0.237) | 14.531<br>(± 0.334) | 14.082<br>(± 0.302) | 14.015<br>(± 0.535) | 12.644<br>(± 0.562) | 6.276<br>(± 0.713) |
| | $T$ | 1.139<br>(± 0.000) | 1.135<br>(± 0.023) | 1.209<br>(± 0.035) | 1.190<br>(± 0.021) | 1.214<br>(± 0.023) | 1.235<br>(± 0.036) | 1.268<br>(± 0.043) | 1.267<br>(± 0.077) |
| | $m$ | 0.704<br>(± 0.000) | 0.750<br>(± 0.095) | 0.644<br>(± 0.021) | 0.654<br>(± 0.028) | 0.716<br>(± 0.049) | 0.529<br>(± 0.062) | 0.568<br>(± 0.093) | 1.761<br>(± 0.247) |
| | $\Theta$ | 128.06<br>(± 0.012) | 128.83<br>(± 1.794) | 125.16<br>(± 1.024) | 125.08<br>(± 1.178) | 123.56<br>(± 1.177) | 123.76<br>(± 1.912) | 121.81<br>(± 1.946) | 128.64<br>(± 4.026) |
| **Charadriiformes** | | | | | | | | | |
| *Numenius phaeopus* | $v_1$ | - | 2.982<br>(± 0.003) | 2.887<br>(± 0.021) | 2.845<br>(± 0.029) | 2.722<br>(± 0.030) | 2.614<br>(± 0.043) | 2.332<br>(± 0.051) | 2.542<br>(± 0.138) |
| | $v_2$ | - | 6.245<br>(± 0.004) | 6.086<br>(± 0.066) | 6.047<br>(± 0.064) | 5.691<br>(± 0.085) | 5.308<br>(± 0.097) | 4.735<br>(± 0.127) | 4.176<br>(± 0.211) |
| | $T$ | - | 1.968<br>(± 0.002) | 1.931<br>(± 0.019) | 1.981<br>(± 0.027) | 1.894<br>(± 0.040) | 1.796<br>(± 0.063) | 1.501<br>(± 0.052) | 2.386<br>(± 0.132) |
| | $m$ | - | 0.056<br>(± 0.000) | 0.055<br>(± 0.001) | 0.056<br>(± 0.001) | 0.052<br>(± 0.003) | 0.042<br>(± 0.004) | 0.023<br>(± 0.007) | 0.133<br>(± 0.013) |
| | $\Theta$ | - | 147.88<br>(± 0.104) | 149.67<br>(± 1.009) | 147.32<br>(± 1.271) | 150.84<br>(± 1.960) | 157.11 (± 3.098) | 173.13<br>(± 3.298) | 141.10<br>(± 6.502) |
| *Tringa brevipes/T. incana* | $v_1$ | 7.894<br>(± 0.135) | 8.487<br>(± 0.093) | 7.516<br>(± 0.166) | 7.014<br>(± 0.223) | 6.382<br>(± 0.267) | 5.258<br>(± 0.625) | 2.806<br>(± 0.292) | 1.016<br>(± 0.086) |

Table 2.2 continued

| | | Parameter | 8:8 | 7:7 | 6:6 | 5:5 | 4:4 | 3:3 | 2:2 | 1:1 |
|---|---|---|---|---|---|---|---|---|---|---|
| | | $v_2$ | 2.559<br>(± 0.045) | 2.835<br>(± 0.036) | 2.663<br>(± 0.055) | 2.537<br>(± 0.085) | 2.613<br>(± 0.103) | 2.395<br>(± 0.111) | 1.416<br>(± 0.150) | 0.578<br>(± 0.050) |
| | | $T$ | 6.575<br>(± 0.134) | 7.624<br>(± 0.107) | 7.284<br>(± 0.189) | 7.153<br>(± 0.291) | 7.542<br>(± 0.364) | 7.033<br>(± 0.389) | 3.856<br>(± 0.536) | 1.942<br>(± 0.203) |
| | | $m$ | 0.0091<br>(± 0.000) | 0.0081<br>(± 0.000) | 0.0084<br>(± 0.000) | 0.0085<br>(± 0.000) | 0.0090<br>(± 0.000) | 0.0098<br>(± 0.000) | 0.008<br>(± 0.002) | 0.165<br>(± 0.015) |
| | | $\Theta$ | 56.345<br>(± 0.986) | 49.828<br>(± 0.628) | 52.707<br>(± 1.250) | 54.627<br>(± 2.022) | 53.799<br>(± 2.686) | 58.978<br>(± 4.510) | 113.161<br>(± 10.291) | 117.030<br>(± 8.657) |
| **Passeriformes** | | | | | | | | | | |
| | *Luscinia svecica* | $v_1$ | - | 3.877<br>(± 0.005) | 3.934<br>(± 0.089) | 4.618<br>(± 0.344) | 5.056<br>(± 0.408) | 5.827<br>(± 0.435) | 6.322<br>(± 0.488) | 5.452<br>(± 0.403) |
| | | $v_2$ | - | 21.452<br>(± 0.092) | 20.980<br>(± 0.442) | 18.847<br>(± 0.961) | 15.954<br>(± 1.072) | 15.795<br>(± 1.156) | 14.675<br>(± 1.307) | 15.969<br>(± 1.432) |
| | | $T$ | - | 1.290<br>(± 0.003) | 1.285<br>(± 0.015) | 1.276<br>(± 0.031) | 1.243<br>(± 0.063) | 1.226<br>(± 0.043) | 1.203<br>(± 0.067) | 1.256<br>(± 0.104) |
| | | $m$ | - | 1.956<br>(± 0.058) | 2.122<br>(± 0.108) | 2.127<br>(± 0.245) | 2.330<br>(± 0.347) | 3.357<br>(± 0.334) | 2.416<br>(± 0.332) | 2.940<br>(± 0.312) |
| | | $\Theta$ | - | 166.94<br>(± 0.176) | 167.608<br>(± 1.033) | 167.935<br>(± 2.405) | 176.558<br>(± 5.360) | 172.299<br>(± 3.636) | 175.525<br>(± 4.166) | 180.675<br>(± 8.488) |
| | *Pinicola enucleator* | $v_1$ | - | 2.519<br>(± 0.016) | 2.846<br>(± 0.057) | 2.843<br>(± 0.076) | 2.658<br>(± 0.113) | 2.597<br>(± 0.120) | 2.197<br>(± 0.121) | 2.325<br>(± 0.117) |
| | | $v_2$ | - | 1.786<br>(± 0.011) | 2.355<br>(± 0.013 | 2.112<br>(± 0.046) | 1.898<br>(± 0.063) | 1.656<br>(± 0.050) | 1.412<br>(± 0.037) | 1.465<br>(± 0.073) |

Table 2.2 continued

| | Parameter | 8:8 | 7:7 | 6:6 | 5:5 | 4:4 | 3:3 | 2:2 | 1:1 |
|---|---|---|---|---|---|---|---|---|---|
| | *T* | - | 1.979<br>(± 0.021) | 2.449<br>(± 0.028) | 2.317<br>(± 0.076) | 2.098<br>(± 0.099) | 1.866<br>(± 0.077) | 1.568<br>(± 0.048) | 2.480<br>(± 0.184) |
| | *m* | - | 0.0073<br>(± 0.001) | 0.0105<br>(± 0.000) | 0.0107<br>(± 0.001) | 0.00677<br>(± 0.001) | 0.0033<br>(± 0.001) | 0.0010<br>(± 0.001) | 0.0596<br>(± 0.004) |
| | *Θ* | - | 223.76<br>(± 1.51) | 197.10<br>(± 1.41) | 205.45<br>(± 3.30) | 219.25<br>(± 5.23) | 233.07<br>(± 5.22) | 256.52<br>(± 4.80) | 212.34<br>(± 11.87) |
| *Pica pica/Pica hudsonia* | $v_1$ | - | 2.699<br>(± 0.042) | 2.485<br>(± 0.046) | 2.406<br>(± 0.057) | 2.298<br>(± 0.075) | 2.300<br>(± 0.094) | 2.117<br>(± 0.142) | 1.567<br>(± 0.144) |
| | $v_2$ | - | 7.107<br>(± 0.126) | 6.759<br>(± 0.225) | 6.470<br>(± 0.330) | 6.604<br>(±0.390) | 6.565<br>(± 0.501) | 5.537<br>(± 0.528) | 3.029<br>(± 0.587) |
| | *T* | - | 3.334<br>(± 0.069) | 3.017<br>(± 0.046) | 2.868<br>(± 0.067) | 2.710<br>(± 0.089) | 2.561<br>(± 0.114) | 2.325<br>(± 0.143) | 2.309<br>(±0.190) |
| | *m* | - | 0.0141<br>(± 0.000) | 0.0121<br>(± 0.000) | 0.0119<br>(± 0.010) | 0.0086<br>(± 0.001) | 0.0066<br>(± 0.001) | 0.0033<br>(± 0.001) | 0.0808<br>(± 0.012) |
| | *Θ* | - | 108.09<br>(± 1.602) | 116.50<br>(± 1.48) | 121.01<br>(± 2.00) | 126.62<br>(± 3.26) | 132.95<br>(± 4.74) | 146.85<br>(± 7.58) | 162.11<br>(± 9.71) |

Table 2.3: Scaled root mean square error (SRMSE) for each parameter at each sample size.

| | Parameter | 7:7 | 6:6 | 5:5 | 4:4 | 3:3 | 2:2 | 1:1 |
|---|---|---|---|---|---|---|---|---|
| **Anseriformes** | | | | | | | | |
| *Clangula hyemalis* | *$v_1$* | - | 0.087 | 0.054 | 0.091 | 0.060 | 0.089 | 0.313 |
| | *$v_2$* | | 0.076 | 0.072 | 0.070 | 0.143 | 0.002 | 0.004 |
| | *T* | - | 0.294 | 0.255 | 0.273 | 0.261 | 0.336 | 0.495 |
| | *m* | - | 0.102 | 0.120 | 0.197 | 0.259 | 0.346 | 0.411 |
| | *Θ* | - | 0.254 | 0.215 | 0.222 | 0.278 | 0.316 | 0.460 |
| *Anas crecca* | | | | | | | | |
| | *$v_1$* | - | - | 0.001 | 0.020 | 0.074 | 0.020 | 0.216 |
| | *$v_2$* | - | - | 0.003 | 0.013 | 0.012 | 0.096 | 0.439 |
| | *T* | - | - | 0.059 | 0.088 | 0.134 | 0.111 | 0.231 |
| | *m* | - | - | 0.116 | 0.113 | 0.053 | 0.559 | 0.038 |
| | *Θ* | - | - | 0.055 | 0.074 | 0.094 | 0.071 | 0.125 |
| *Anas penelope/A. americana* | *$v_1$* | 0.038 | 0.027 | 0.005 | 0.021 | 0.076 | 0.007 | 0.072 |
| | *$v_2$* | 0.030 | 0.048 | 0.074 | 0.108 | 0.111 | 0.234 | 1.487 |
| | *T* | 0.004 | 0.058 | 0.042 | 0.061 | 0.077 | 0.101 | 0.101 |
| | *m* | 0.061 | 0.093 | 0.076 | 0.017 | 0.331 | 0.239 | 0.600 |
| | *Θ* | 0.006 | 0.023 | 0.024 | 0.036 | 0.035 | 0.051 | 0.004 |
| **Charadriiformes** | | | | | | | | |
| *Numenius phaeopus* | *$v_1$* | - | 0.032 | 0.047 | 0.094 | 0.140 | 0.277 | 0.172 |
| | *$v_2$* | - | 0.026 | 0.032 | 0.097 | 0.176 | 0.318 | 0.495 |
| | *T* | - | 0.020 | 0.006 | 0.040 | 0.096 | 0.308 | 0.174 |
| | *m* | - | 0.018 | 0.018 | 0.096 | 0.357 | 1.435 | 0.579 |
| | *Θ* | - | 0.012 | 0.004 | 0.019 | 0.058 | 0.146 | 0.048 |

Table 2.3 continued

| | Parameter | 7:7 | 6:6 | 5:5 | 4:4 | 3:3 | 2:2 | 1:1 |
|---|---|---|---|---|---|---|---|---|
| | $v_2$ | 0.090 | 0.031 | 0.017 | 0.013 | 0.077 | 0.821 | 3.464 |
| | $T$ | 0.129 | 0.089 | 0.072 | 0.120 | 0.056 | 0.721 | 2.417 |
| | $m$ | 0.123 | 0.119 | 0.706 | 0.006 | 0.102 | 0.125 | 0.945 |
| | $\Theta$ | 0.139 | 0.077 | 0.039 | 0.055 | 0.038 | 0.498 | 1.028 |
| **Passeriformes** | | | | | | | | |
| *Luscinia svecica* | $v_1$ | - | 0.113 | 0.399 | 0.460 | 0.495 | 0.541 | 0.463 |
| | $v_2$ | - | 0.106 | 0.285 | 0.479 | 0.509 | 0.641 | 0.563 |
| | $T$ | - | 0.056 | 0.118 | 0.249 | 0.177 | 0.286 | 0.407 |
| | $m$ | - | 0.262 | 0.570 | 0.745 | 0.641 | 0.700 | 0.619 |
| | $\Theta$ | - | 0.030 | 0.070 | 0.159 | 0.109 | 0.126 | 0.242 |
| *Pinicola enucleator* | $v_1$ | - | 0.117 | 0.116 | 0.055 | 0.032 | 0.144 | 0.080 |
| | $v_2$ | - | 0.241 | 0.154 | 0.058 | 0.078 | 0.265 | 0.219 |
| | $T$ | - | 0.194 | 0.148 | 0.059 | 0.058 | 0.259 | 0.204 |
| | $m$ | - | 0.286 | 0.280 | 0.148 | 1.212 | 7.000 | 0.872 |
| | $\Theta$ | - | 0.143 | 0.096 | 0.027 | 0.033 | 0.122 | 0.061 |
| *Pica pica/Pica hudsonia* | $v_1$ | - | 0.083 | 0.119 | 0.171 | 0.170 | 0.272 | 0.718 |
| | $v_2$ | - | 0.042 | 0.089 | 0.067 | 0.073 | 0.272 | 1.325 |
| | $T$ | - | 0.089 | 0.146 | 0.213 | 0.283 | 0.414 | 0.424 |
| | $m$ | - | 0.165 | 0.168 | 0.581 | 1.061 | 3.333 | 0.829 |
| | $\Theta$ | - | 0.088 | 0.122 | 0.161 | 0.201 | 0.276 | 0.344 |

## 2.8 References

Afgan, E, Baker D, van den Beek M, *et al.* (2016) The Galaxy platform for accessible, reproducible, and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, **44**, W3-W10.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics, 30, 2114-2120.

Danecek P, Auton A, Abecasis G, *et al*. (2011) The Variant Call Format and VCFtools. *Bioinformatic*s, **27**, 2156-2158.

Faircloth BC (2013) illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. http://dx.doi.org/10.6079/J9ILL.

Faircloth BC (2016) PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics,* **32**,786-788.

Felsenstein J (2005) Accuracy of coalescent likelihood estimates: do we need more sites, more sequences, or more loci? *Molecular Biology and Evolution*, **23**, 691-700.

Flaxman SM, Wacholder AC, Feder JL, Nosil P (2014) Theoretical models of the influence of genomic architecture on the dynamics of speciation. *Molecular Ecology*, **23**, 4074-4088.

Glenn TC, Nilsen R, Kieran TJ, *et al.* (2016) Adapterama I: Universal stubs and primers for thousands of dual-indexed Illumina libraries (iTru & iNext). Accepted at Molecular Ecology Resources, pending minor revisions, available at http://biorxiv.org/content/early/2016/06/15/049114

Grabherr MG, Haas BJ, Yassour M, *et al.* (2011) Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature Biotechnology*, **29**, 644-652.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLOS Genetics,* **5**, e1000695.

Humphries E, Winker K (2011) Discord reigns among nuclear, mitochondrial, and phenotypic estimates of divergence in nine lineages of Beringian birds. *Molecular Ecology,* **20**, 573-583.

Jeffries DL, Copp GH, Lawson Handley L, Olsén KH, Sayer CD, Hanfling B (2016) Comparing RADsxeq and microsatellites to infer complex phylogeographic patterns, an empirical perspective in the Crucian carp, *Carassius carassius*, L. *Molecular Ecology*, **25**, 2997-3018.

Kalinowski ST (2005) Do polymorphic loci require large sample sizes to estimate genetic distances? *Heredity*, **94**, 33-36.

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN]

Li H, Durbin R (2009) Fats and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25, 1754-1760.

Li H, Handsaker B, Wysoker A, *et al*. (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics,* **25**, 2078-2079.

McKenna A, Hanna M, Banks E, *et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research,* **20**, 1297-1303.

Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, **9**, 66-73.

Nazareno AG, Jump AS (2012) Species-genetic diversity correlations in habitat fragmentation can be biased by small sample sizes. *Molecular Ecology*, **21**, 2847-2849.

Nazareno AG, Bemmels JB, Dick CW, Lohmann LG (2017) Minimum sample sizes for population genomics: an empirical study from an Amazonian plant species. *Molecular Ecology Resource* accepted article; doi:10.1111/1755-0998.12654.

Nosil P, Feder JL, Flaxman SM, Gompert Z (2017) Tipping points in the dynamics of speciation. *Nature Ecology and Evolution*, **1**, 0001.

Pruett CL, Winker K (2008) The effects of sample size on population genetic diversity estimates in song sparrows *Melospiza melodia*. *Journal of Avian Biology*, **39**, 252-256.

Robinson JD, Coffman AJ, Hickerson MJ, Gutenkunst RN (2014) Sampling strategies for frequency spectrum-based population genomic inference. *BMC Evolutionary Biology*, **14**, 254-270.

Willing E-M, Dreyer C, van Oosterhout C (2012) Estimates of genetic differentiation measured by $F_{ST}$ do not necessarily require large sample sizes when using many SNP markers. *PLOS One*, **7**, e42649.

## 2.9 Appendix to Chapter 2: Supplementary Figures

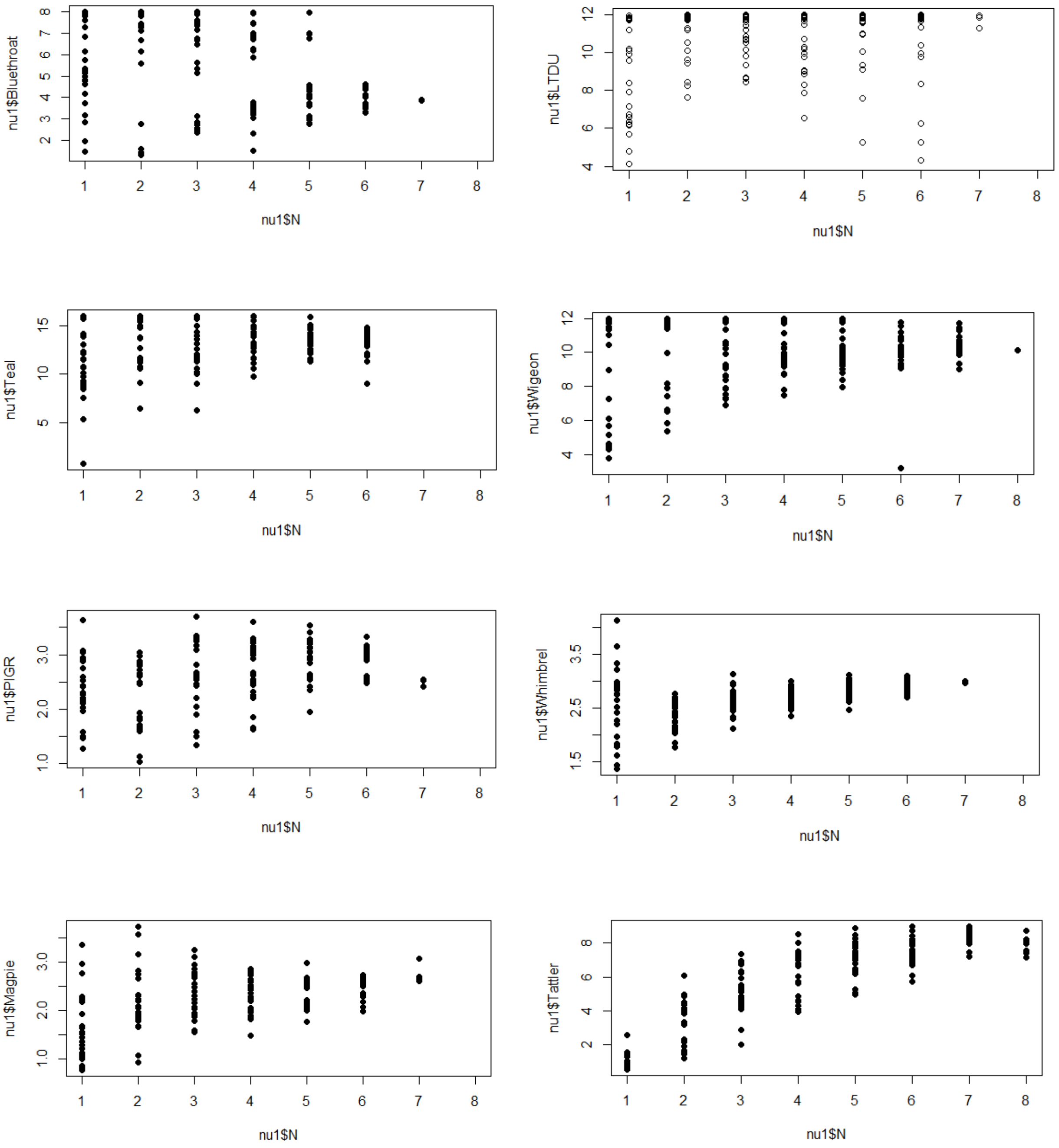

Figure S1: Estimates of $v_1$ at varying sample sizes in eight lineages.

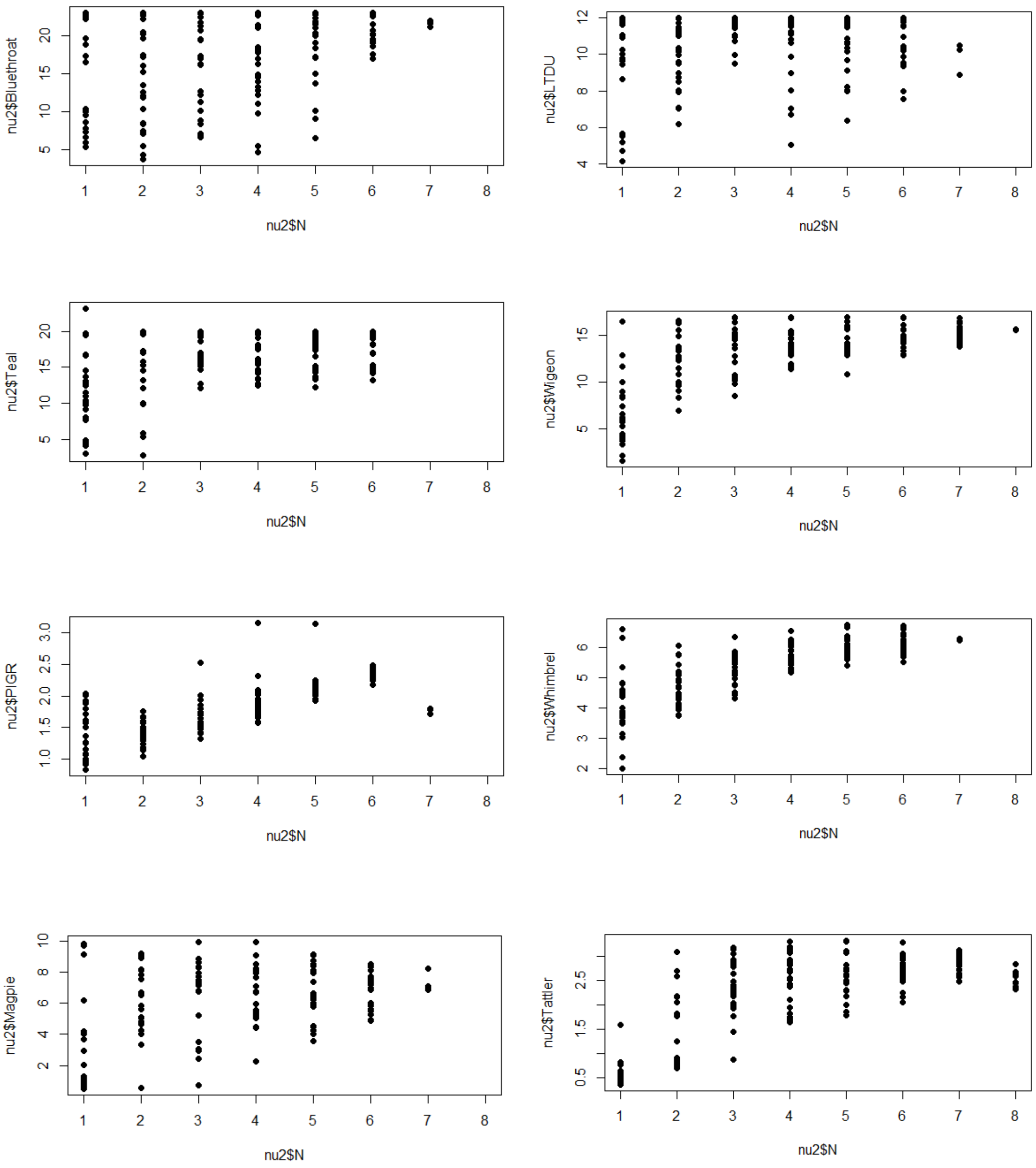

Figure S2: Estimates of $v_2$ at varying sample sizes in eight lineages.

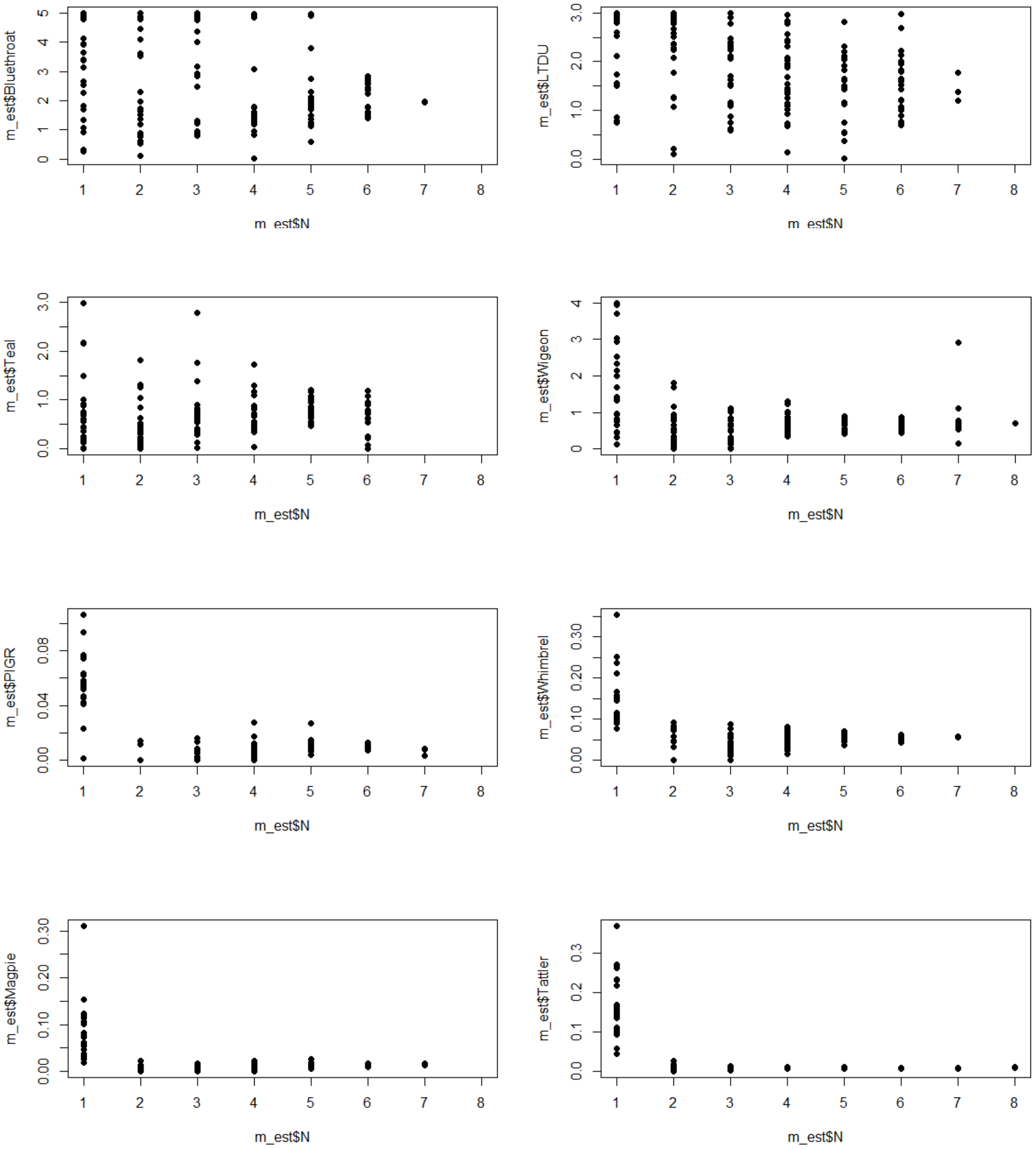

Figure S3: Estimates of *m* at varying sample sizes in eight lineages.

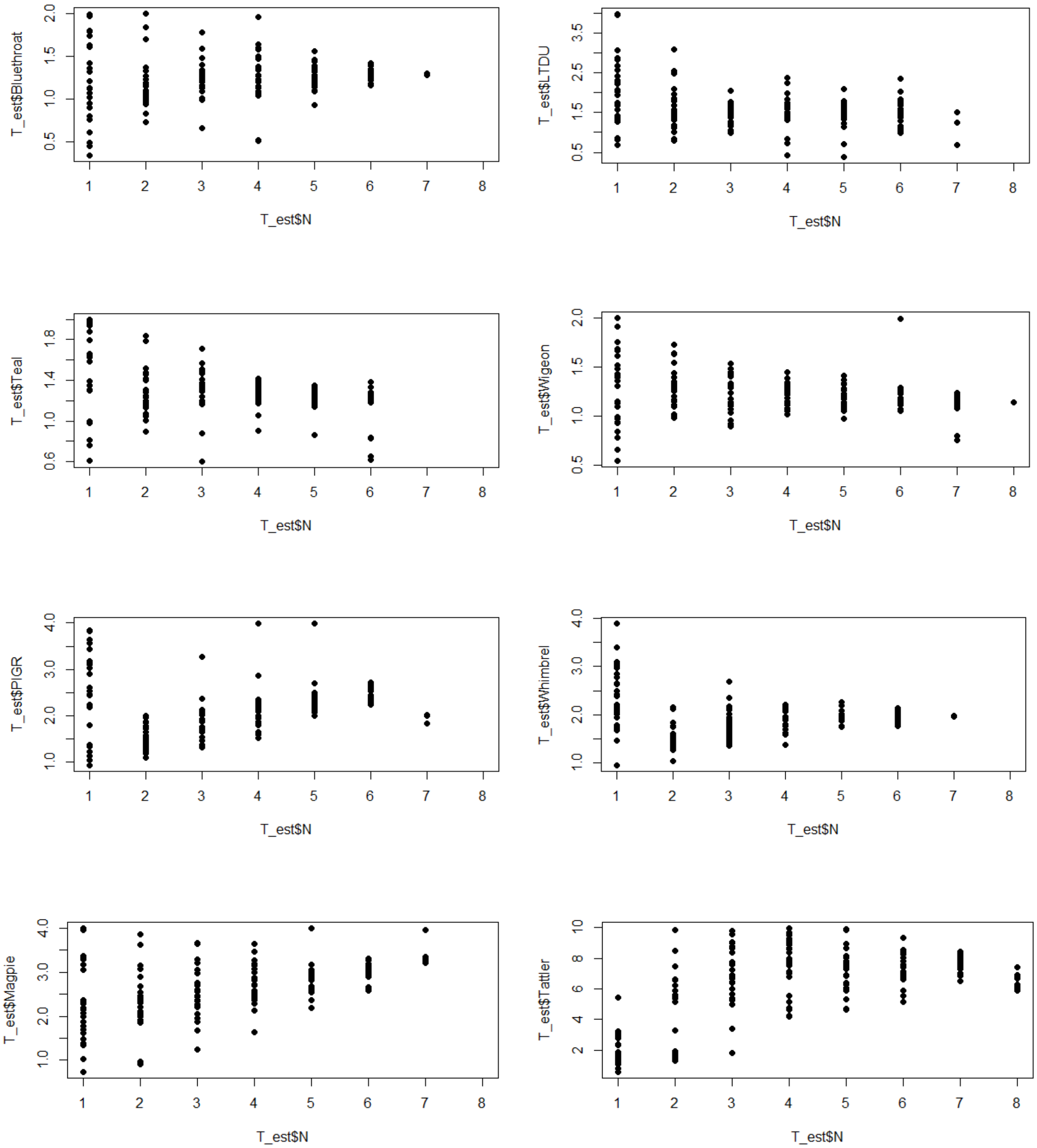

Figure S4: Estimates of *T* at varying sample sizes in eight lineages.

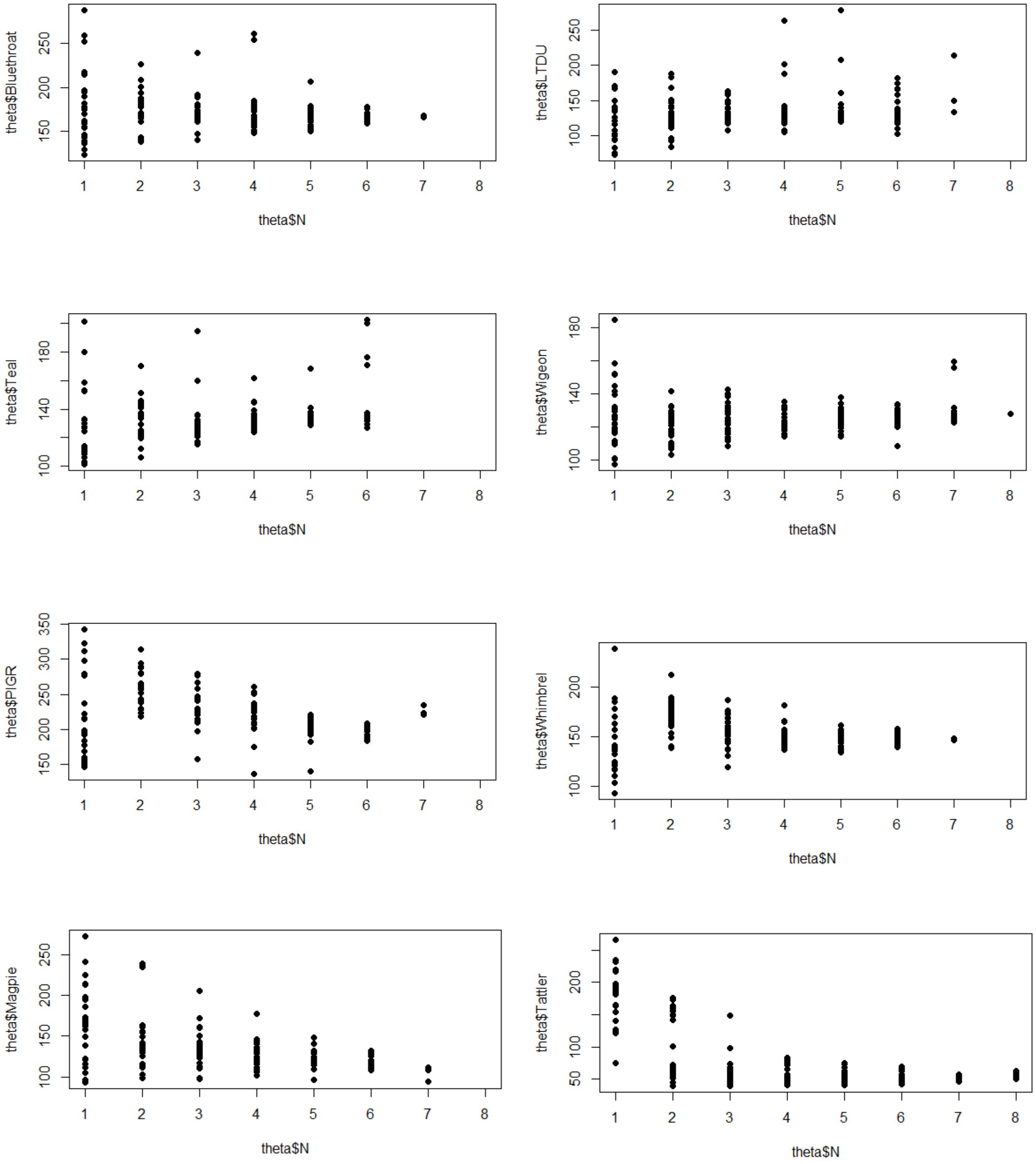

Figure S5: Estimates of *Θ* at varying sample sizes in eight lineages.

## Conclusions

Understanding processes of divergence and how the speciation continuum develops is key to understanding the production and maintenance of biodiversity. Despite the diversity of these lineages, the divergence history was not best explained by different models.  In all of the eight two-population splits investigated, all best fit a split-migration, speciation with gene flow model. Additionally, estimates of divergence ($F_{ST}$) and gene flow clustered into two distinct groups: a lower divergence, higher gene flow group; and a relatively high divergence, low gene flow group, a situation that has been reported previously in the speciation literature (Flaxman *et al.* 2014, Nosil *et al.* 2017). Whether this reflects broad trends in speciation among Beringian birds will not be clear until more lineages are sampled.

Population demographics such as those made here are dependent on adequate sample sizes. For some estimates, particularly gene flow and effective population sizes, parameters were reliably over- or underestimated at low sample sizes, and estimates of gene flow were particularly robust with few individuals sampled per population. However, the accuracy of estimates in population-level splits decreased more quickly than in other lineages, and if possible additional individuals should be sampled in such situations, with at least 6 individuals per population representing a potential sampling goal.

## General Appendix: List of Supplementary Data Files

All files archived as supplemental files.

Reference sequences:

bluethroat_refseq.fasta

grosbeak_refseq.fasta

longtail_refseq.fasta

magpie_refseq.fasta

plover_refseq.fasta

tattler_refseq.fasta

teal_refseq.fasta

whimbrel_refseq.fasta

wigeon_refseq.fasta

All called SNPs:

bluethroat_allSNPs.vcf

grosbeak_allSNPs.vcf

longtail_allSNPs.vcf

magpie_allSNPs.vcf

plover_allSNPs.vcf

tattler_allSNPs.vcf

teal_allSNPs.vcf

whimbrel_allSNPs.vcf

wigeon_allSNPs.vcf

Thinned (one SNP per locus), biallelic SNPs, with Z-linked loci removed:

bluethroat_thinnedSNPs.vcf

grosbeak_thinnedSNPs.vcf

longtail_thinnedSNPs.vcf

magpie_thinnedSNPs.vcf

plover_thinnedSNPs.vcf

tattler_thinnedSNPs.vcf

teal_thinnedSNPs.vcf

whimbrel_thinnedSNPs.vcf

wigeon_thinnedSNPs.vcf