USGS
science for a changing world

Prepared in cooperation with the California Department of Fish and Wildlife

Identification of Single Nucleotide Polymorphisms for Use in a Genetic Stock Identification System for Greater White-Fronted Goose (*Anser albifrons*) Subspecies Wintering in California

Open-File Report 2019-1040

U.S. Department of the Interior
U.S. Geological Survey

Cover: Tule white-fronted goose in the Kahiltna Valley, Alaska. Photograph by Craig Ely, U.S. Geological Survey, 1984.

Identification of Single Nucleotide Polymorphisms for Use in a Genetic Stock Identification System for Greater White-Fronted Goose (*Anser albifrons*) Subspecies Wintering in California

By Robert E. Wilson, Sarah A. Sonsthagen, Jeffrey M. DaCosta, Craig R. Ely, Michael D. Sorenson, and Sandra L. Talbot

Prepared in cooperation with the California Department of Fish and Wildlife

Open-File Report 2019–1040

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2019

Suggested citation:
Wilson, R.E., Sonsthagen, S.A., DaCosta, J.M., Ely, C.R., Sorenson, M.D., and Talbot, S.L., 2019, Identification of single nucleotide polymorphisms for use in a genetic stock identification system for greater white-fronted goose (*Anser albifrons*) subspecies wintering in California: U.S. Geological Survey Open-File Report 2019-1040, 18 p., https://doi.org/10.3133/ofr20191040.

ISSN 2331-1258 (online)

Contents

Figures

Tables

Conversion Factors

International System of Units to U.S. customary units

Multiply	By	To obtain
	Mass	
microgram (μg)	0.00000003527	ounce, avoirdupois (oz)
	Deoxyribonucleic acid concentration	
nanogram per microliter (ng/μL)	0.000008345404	pounds per gallon

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as follows:

°F = (1.8 × °C) + 32.

Abbreviations

A	the nucleobase adenine
BBL	Bristol Bay Lowlands
C	the nucleobase cytosine
CIB	Cook Inlet Basin
ddRAD-seq	double-digest RAD seq
DNA	deoxyribonucleic acid
G	the nucleobase guanine
GWFG	greater white-fronted geese
MCMC	Markov Chain Monte Carlo
mtDNA	mitochondrial DNA
PCA	Principal Components Analysis
PCR	polymerase chain reaction
RADseq	restriction site-associated DNA sequencing
SNP	single nucleotide polymorphism
STA	Specific-Target Amplification
T	the nucleobase thymine
YKD	Yukon-Kuskokwim Delta

Identification of Single Nucleotide Polymorphisms for Use in a Genetic Stock Identification System for Greater White-Fronted Goose (*Anser albifrons*) Subspecies Wintering in California

By Robert E. Wilson[1], Sarah A. Sonsthagen[1], Jeffrey M. DaCosta[2], Craig R. Ely[1], Michael D. Sorenson[3], and Sandra L. Talbot[1]

Abstract

California provides wintering habitat for most greater white-fronted geese (*Anser albifrons* [GWFG]) in the Pacific Flyway and this population has rapidly increased since the 1980s. Increased harvest of GWFG wintering in California may prevent agricultural depredation while providing increased hunting opportunities. However, changes in harvest levels are unlikely to be uniform across the species because of the presence of multiple subspecies of GWFG in the Pacific Flyway, each with their own population distribution and trends. White-fronted geese in the Cook Inlet Basin of south-central Alaska, a potentially vulnerable subspecies (Tule goose, *A. a. elgasi*), are among the geese that winter predominantly in the Sacramento Valley and Suisun and Napa marshes of north-central California. Efforts to limit sport harvest of Tule geese are complicated because although the subspecies is phenotypically larger and darker in color than other subspecies, they can be difficult to identify in the field and in hunter bag checks. To assist in an accurate assessment of Tule goose harvest, we used double-digest restriction site-associated deoxyribonucleic acid sequencing (ddRAD-seq) techniques to develop a genetic stock identification panel of single nucleotide polymorphisms (SNPs) to differentiate Tule geese from individuals belonging to other GWFG subspecies and populations that winter in California. Although the panel we developed was designed and tested for Fluidigm SNP-type technology, the ddRAD-seq sequences can be used to design SNP panels for use in other platforms.

[1] U.S. Geological Survey Alaska Science Center.
[2] Boston College.
[3] Boston University.

Introduction

Greater white-fronted geese (*Anser albifrons* [GWFG]), a migratory species with a Holarctic distribution, are harvested by sport and subsistence hunters throughout much of their range, especially in North America. Based mainly on band recovery data, North American GWFG have been delineated into two major populations: (1) Pacific Flyway, and (2) Midcontinent (Central Flyway). California provides wintering habitat for most GWFG using the Pacific Flyway. The wintering population in California has rapidly increased from fewer than 100,000 in the early 1980s to more than 600,000 more recently (Olson, 2014). This increase has led to a proposed increase in hunter harvest to reduce agricultural damage by geese while providing increased hunting opportunities for sport and subsistence users (U.S. Fish and Wildlife Service, 2014). However, changes in wintering ground harvest levels are unlikely to be uniform across the species, owing to the co-occurrence of three nesting populations of Pacific Flyway GWFG on wintering areas of California, each characterized by different population distributions and trends. Although more than 90 percent of Pacific Flyway GWFG breed on the Yukon-Kuskokwim Delta (YKD) in western Alaska, breeding populations also are present in the Bristol Bay Lowlands (BBL) of southwestern Alaska and the Cook Inlet Basin (CIB) of south-central Alaska (Ely and Dzubin, 1994). Previous investigations of GWFG in the Pacific Flyway reported differences among these three breeding populations with respect to morphology (Orthmeyer and others, 1995; Ely and others, 2005), distribution (Ely and Takekawa, 1996), timing of migration and reproduction (Ely and Takekawa, 1996; Ely, 2008), and genetics (Ely and others, 2017; Wilson and others, 2018). The three Pacific Flyway nesting populations are allopatric during the summer nesting season, but overlap in distribution during the non-breeding season (Ely and Takekawa, 1996; Ely, 2008; see fig. 1). Wildlife managers are particularly interested in the CIB population, which is recognized as a distinct subspecies (Tule goose, *A. a. elgasi*) that winters predominantly near Sacramento, Delevan, and Colusa National Wildlife Refuges in the Sacramento Valley and Suisun and Napa marshes of north-central California (Deuel and Takekawa, 2008).

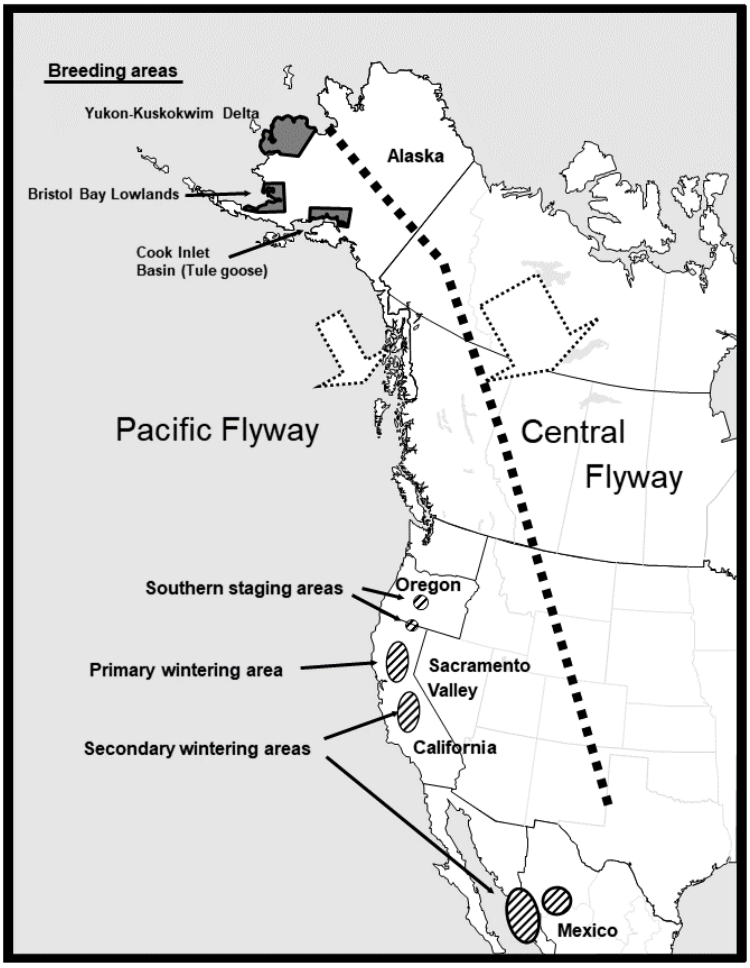

Figure 1. Image showing distribution and area use of greater white-fronted geese in the Pacific Flyway. Black dashed line indicates the proposed boundary between the Pacific and Central flyways. Large dotted arrows indicate direction of winter migration. Shaded areas in Alaska indicate the three main nesting areas. Circles with diagonal lines indicate main areas used for staging and winter.

Because the Tule goose population is small (likely less than [<]10,000 birds; Deuel and Takekawa, 2008), and perhaps vulnerable, ongoing efforts have been made to curtail harvest of these birds by shortening season lengths and bag limits near their primary wintering areas (Mensik, 1991). Efforts to limit sport harvest of Tule geese through use of physical features (Tule geese are larger and darker in color than BBL and YKD geese; Orthmeyer and others, 1995) are hampered by imperfect identification in the field and in hunter bag checks. Current harvest estimates of Tule geese on National Wildlife Refuges in the Sacramento Valley are estimated at <500 birds per season (Mensik, 1991; U.S. Fish and Wildlife Service, 2014), but the accuracy of these estimates is unknown given the overlap in size of Tule geese and GWFG from other populations (Orthmeyer and others, 1995). Additionally, proposed changes to overall harvest of GWFG in the Pacific Flyway (U.S. Fish and Wildlife Service, 2014) could also increase the harvest of Tule geese.

To assist in an accurate assessment of Tule goose harvest, we developed a genetic stock identification system that will assist in providing reliable harvest composition estimates of GWFG in the Pacific Flyway. In our prior research, we reported significant levels of population differentiation between Tule geese and other GWFG populations (Ely and others, 2017; Wilson and others, 2018). However, exploratory genetic assignment tests based on fragment data from 10 nuclear microsatellite loci and sequence data from the mitochondrial deoxyribonucleic acid (mtDNA) control region, done using the assignment calculator MLE 1.0 (available at https://msu.edu/~scribne3/lab/programs.htm), indicated a misassignment rate of about 13 percent between non-Tule Pacific Flyway populations, one representative Midcontinent population and Tule geese. Thus, we concluded that additional loci are required to accurately assign GWFG to their population of origin. Although microsatellite markers traditionally have been the marker of choice for genetic stock identification, reduced-representation genome sequencing technologies, including restriction site-associated DNA sequencing (RADseq), recently have facilitated the development of single nucleotide polymorphism (SNP) panels in non-model organisms (Contina and others, 2017; Bay and others, 2018). Therefore, we used a double-digest RADseq (ddRAD-seq) protocol to detect genome-wide SNPs from an ascertainment panel of 239 samples including representatives of GWFG from the entire circumpolar range. The goal of the ddRAD-seq effort was to identify SNPs that are highly differentiated between the Tule goose and other GWFG populations, and ultimately to develop a downscaled multi-locus SNP panel that can be used to determine harvest composition of GWFG in California. The multi-locus SNP panel is easily adapted for use in different laboratories, which will facilitate the development of management-related strategies particularly within the context of monitoring the harvest of the Tule goose from non-nesting aggregations of GWFG wintering in California.

Methods

Sampling and Deoxyribonucleic Acid Extraction

Blood was sampled from 239 GWFG from across the global range of the species (see table 1 for sample sizes and localities; Wilson and others 2018) for the ddRAD-seq dataset. Genomic deoxyribonucleic acid (DNA) was extracted using a DNeasy® Blood & Tissue Kit and following manufacturer protocols (Qiagen, Valencia, California). Extractions were quantified using a Modulus™ Microplate (Turner BioSystems, Inc., Sunnyvale, California) and a Broad Range Quant-iT™ dsDNA Assay Kit (Thermo Fisher Scientific, Inc., Waltham, Massachusetts) to ensure a minimum DNA concentration of 10 ng/µL.

Table 1. Location and sample size information for greater white-fronted geese examined in this study.

[N, sample size]

Continent	Flyway	Locality	N
North America	Pacific	Cook Inlet (Tule), Alaska	25
North America	Pacific	Bristol Bay, Alaska	17
North America	Pacific	Yukon Kuskokwim-Delta, Alaska	23
North America	Midcontinent	Interior Alaska	20
North America	Midcontinent	Arctic Coastal Plain, Alaska	20
North America	Midcontinent	Canada Arctic	40
North America	Midcontinent	Western Interior Canada—Old Crow Flats	20
Asia	Eastern Palearctic	Anadyr, Kolyma, Magadan, Russia	24
Asia	Western Palearctic	Lena River—Taimyr, Russia	30
Greenland	Western Palearctic	Greenland	20

Preparation of Double-Digest Restriction Site-Associated Deoxyribonucleic Acid Sequencing Library

Sample preparation for ddRAD sequencing followed the double-digest protocol outlined in DaCosta and Sorenson (2014). Genomic DNA (about 1 µg) was digested with high-fidelity versions of *Sbf*I and *Eco*RI restriction enzymes (New England Biolabs, Ipswich, Massachusetts). Amplification and sequencing adapters containing unique barcode or index sequences were ligated to the sticky ends generated by the restriction enzymes. The samples were then run on 2-percent low-melt agarose gels and DNA fragments ranging from 300 to 450 base pairs (bp) in length (178 to 328 bp in length, excluding adapters) were selected. DNA was extracted from the gel using a MinElute® Gel Extraction Kit (Qiagen) following manufacturer protocol. Size-selected fragments were then amplified via the polymerase chain reaction (PCR) using Phusion™ High-Fidelity DNA Polymerase (Thermo Fisher Scientific) for 20 cycles, and the amplified products were purified using magnetic AMPure® XP beads (Beckman Coulter, Inc., Indianapolis, Indiana). Quantitative PCR with a KAPA™ Library Quantification Kit for Illumina® sequencing platforms (KAPA Biosystems, Wilmington, Massachusetts) was used to quantify the concentration of purified PCR products, and samples were pooled in equimolar concentrations. A multiplexed library was sequenced as a single-end, 150-bp run on an Illumina® HiSeq 2500 at the Tufts University Core Genomics Facility.

Bioinformatics

Raw Illumina reads were processed using a computational pipeline described by DaCosta and Sorenson (2014). First, reads were assigned to individual samples based on barcode/index sequences using bcl2fastq-1.8.4 software (Illumina Inc., San Diego, California) by the Tufts University Core Genomics facility. Next, pre-processing of reads was done using a custom python script (*ddRAD_fastq_qc.py*, Jeffrey DaCosta, Boston College, unpub. data, 2017) that added a "CC" at the beginning of each sequence to complete the *Sbf*I recognition site and removed chimera sequences (that is, sequences that contained complete *Sbf*I or *Eco*RI sites, which result from ligation of separate loci) or reads containing two or more mismatches in the *Sbf*I recognition site. The script also removed the adapter sequence from reads with loci that

were shorter than the read length. For each sample, identical reads were collapsed (while maintaining read counts) using the *CondenseSequences.py* script, and low-quality reads (that is, sequences that failed to cluster with any other reads [−id setting of 0.90] and an average per-base Phred score <20 were filtered out with the *FilterSequences.py* script and the UCLUST function in USEARCH v.5 (Edgar, 2010). Condensed and filtered reads from all samples were then concatenated and clustered with an −id setting of 0.85 in UCLUST. A representative sequence from each cluster was then aligned to the *Gallus gallus* (GenBank assembly reference GCA_000002315.2) reference genome using blastn v.2 (Altschul and others, 1990), and clusters with hits to the same genomic region were identified and joined using the *CombineClusters.py* script. MUSCLE v.3 (Edgar, 2004) was used to align the reads within each cluster, and individual sample genotypes were called using the *RADGenotypes.py* script. Homozygotes and heterozygotes were identified based on thresholds outlined in DaCosta and Sorenson (2014), with individual genotypes assigned to four categories: (1) "missing" (no data), (2) "good" (unambiguously genotyped), (3) "low depth" (<5 reads), and (4) "flagged" (recovered heterozygous genotype, but with haplotype counts outside of acceptable thresholds or with >2 alleles detected). We used Geneious v.10 (Biomaters Inc. San Francisco, California) to manually check and edit a subset of loci flagged as potentially problematic by the genotyping code. To limit any biases due to sequencing error and (or) allelic dropout, alleles with less than 5× coverage were scored as missing, such that a minimum of 10 total reads was required to score a genotype as heterozygous. Polymorphic loci with a median depth of 10, <10 percent missing genotypes, and <10 percent flagged genotypes (of 239 individuals total) were retained for downstream analyses.

Population Structure and Single Nucleotide Polymorphism Discovery

Pairwise phi-st (variance in allele frequencies between groups) and nucleotide diversity for each ddRAD-seq locus and all loci combined were calculated using a custom Python script obtained from J. M. DaCosta (available at https://github.com/BU-RAD-seq/Out-Conversions). Additionally, we did a Principal Components Analysis (PCA) following the methodology of Novembre and Stephens (2008) to explore the genetic clustering of geographic regions. This analysis used all bi-allelic SNPs that were extracted with a custom Python script and the subsequent PCA was done using the prcomp function in R (https://www.r-project.org) to identify nuclear loci/SNPs with the potential to differentiate Tule goose from other GWFG North American populations. Additionally, we completed a PCA using haplotype data using the "dudi.pca" function in the R package *adegenet* (Dray and Dufour, 2007; Jombart 2008).

Loci showing an elevated level of divergence (phi-st >0.1) as well as SNPs with higher contributions to the first PC axis (that is, candidate loci for differentiating the Tule goose population) were further evaluated for inclusion in a 96-locus SNP Fluidigm Corporation EP1™ Genotyping System panel. Following recommendations for locus selection and design of Fluidigm SNP-type assays, SNPs were selected based on the following criteria:

1. Sequences containing target SNPs should have at least 60 bp on either side of the SNP,
2. Only one target SNP is identified per sequence,
3. No insertions or deletions >10 bp are allowed in the sequence,
4. No non-biallelic SNPs are allowed,
5. No adjacent (secondary) SNPs within 30 bp of at least one side of the target SNP are allowed, and
6. guanine-cytosine (GC) content is < 65 percent.

First, adjacent SNPs with a minor allele frequency of 1 percent were excluded. Next, if assay design failed, we used a minor allele frequency threshold of 2 percent. To reduce the potential impact of allelic dropout due to the presence of secondary SNPs, we only selected targeted SNPs with no adjacent SNPs within 10 bp on at least one side. For highly diagnostic loci for which flanking sequences were not sufficient for inclusion in a Fluidigm SNP panel (that is, there were < 60 bp on either side of the target SNP), we used alignments of sequences from the pink-footed goose (*Anser brachyrhynchus*, Genbank Assembly accession GCA_002592135.1) and swan goose (*Anser cygnoides*, Genbank Assembly accession GCA_002166845.1) genomes to estimate likely flanking sequences. Flanking regions were compared across these genomes and, if identical or containing only 1 polymorphic site, the flanking sequence of the pink-footed goose was used to increase the length of the locus. To augment 93 nuclear loci identified for inclusion in the SNP panel, we also included 3 SNPs identified from the mtDNA control region, leveraged from a previously published dataset (Ely and others, 2017; Wilson and others, 2018). These three SNPs showed frequency differences between Tule goose and either Pacific or Midcontinent GWFG populations (fig. 2).

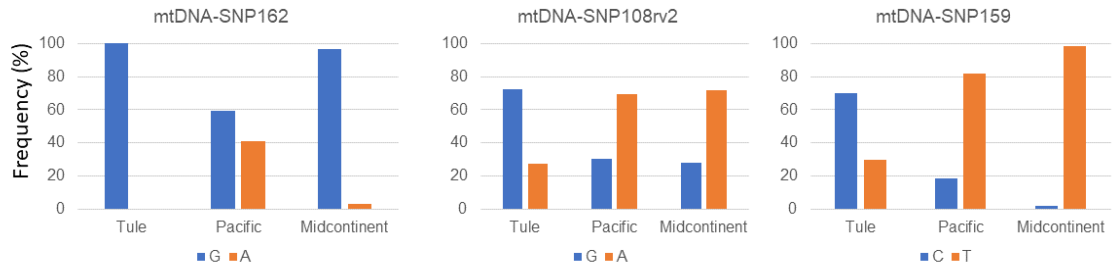

Figure 2. Graphs showing single nucleotide polymorphism (SNP) frequencies in Tule goose, Pacific (non-Tule), and Midcontinent greater white-fronted geese populations for three mitochondrial deoxyribonucleic acid (mtDNA) control region SNPs chosen for inclusion in SNP panel. G, the nucleobase guanine; A, the nucleobase adenine; C, the nucleobase cytosine; T, the nucleobase thymine.

To verify that the subspecies structure estimated from the 3,888 loci in the ddRAD-seq dataset (fig. 3A) was retained in the targeted 93 nuclear autosomal SNPs, we did a PCA implemented in the *adegenet* R package (that is, "dudi.pca") using the SNP calls from ddRAD-seq data and plotted the first two principal components. Additionally, we used STRUCTURE 2.2.3 (Pritchard and others 2000) to determine the level of genetic structure. STRUCTURE assigns individuals to populations maximizing Hardy-Weinberg expectations and minimizing linkage disequilibrium. The analysis was run for K=2, where K is the number of populations, using an admixture model with 100,000 burn-in and 1,000,000 Markov Chain Monte Carlo (MCMC) iterations. The analysis was repeated five times.

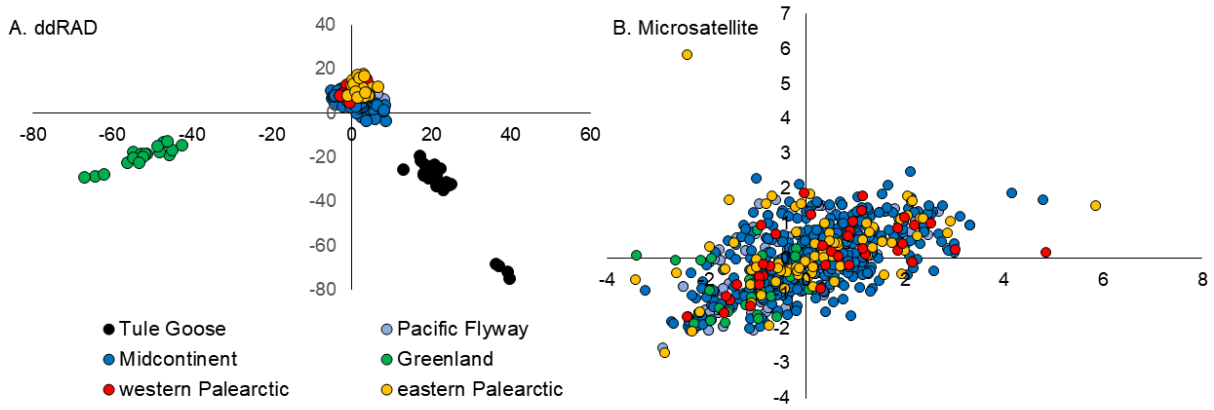

Figure 3. Scatterplot showing (A) first two principal components based on 3,888 double-digest restriction site-associated deoxyribonucleic acid sequencing (ddRAD-seq) loci, and (B) eight microsatellite loci of circumpolar greater white-fronted goose samples. Same pattern is observed if Principal Components Analysis (PCA) is done on haplotypes or with all bi-allelic allelic single nucleotide polymorphisms for ddRAD-seq loci. Pacific Flyway refers to non-Tule nesting populations of greater white-fronted geese in Bristol Bay Lowlands and Yukon-Kuskokwim Delta, Alaska. Numbers on x- and y-axes are PCA coordinates and represent values associated with PC1 (x-axis) and PC2 (y-axis).

Validation of the 96-Single Nucleotide Polymorphism Panel

The Fluidigm Corporation EP1™ Genotyping System was used to genotype the 96-SNP loci (see Wilson and others, 2019 for SNP assay design), using 95 individuals per run and 1 non-template control per genotyping run, in 2 genotyping runs. The goals of this screening of the 96-SNP locus panel were to:

1. Verify that the SNP primers amplified a product;
2. Verify that the SNPs amplified in the panel were sufficient to distinguish Tule geese from the other Pacific Flyway GWFG populations that winter in California; and
3. Assess the level, if any, of allelic dropout in DNA samples sourced from feathers by amplifying tongue and feather samples from twenty-two samples.

Two genotyping runs were completed:

- **Run 1.** This run comprised 95 samples, including (1) DNA extracted from 47 tongue tissue (see section, "Sampling and DNA Extraction") from hunter-harvested geese during the 2014–15 hunting season assigned to subspecies based on phenotype at Delevan National Wildlife Refuge and Sacramento National Wildlife Refuge hunter-check stations, and 48 samples extracted from blood tissue from known breeding areas of Tule geese (n = 28) and non-Tule (n = 20, BBL and YKD). Eleven of the blood samples were used in the reference ddRAD-seq dataset (Tule n =5 and non-Tule n = 6). We included these 11 samples to verify that Fluidigm SNP calls could differentiate Tule and non-Tule geese. This run included a pre-amplification step using Locus-Specific Primers (LSP) and Specific-Target Amplification (STA) primers.
- **Run 2.** This run comprised 95 samples, including (1) 47 samples of DNA extracted from feathers assigned to subspecies based on phenotype collected at hunter-check stations, and (2) 48 samples extracted from blood tissue from known breeders (Tule and non-Tule

geese from the Pacific Flyway). The blood samples were the same as in Run 1, but were not pre-amplified to determine if this extra step was needed.

Fluidigm recommends an initial pre-amplification step for samples with low-quality or low-concentration DNA, as would be expected from feathers. This pre-amplification step was completed using a primer pool containing 96 unlabeled LSP and STA primers. The STA is a multiplex PCR reaction using 50 nanomoles of each primer to amplify the targeted locus using PCR conditions of 95 °C for 15 min and 14 cycles of 95 °C for 15 sec followed by 60 °C for 4 min. PCR products were diluted 1:100 and re-amplified using fluorescently labeled allele-specific and locus-specific primers. The results were imaged on an EP1 Array Reader and alleles were called using Fluidigm automated Genotyping Analysis Software (Fluidigm Inc.) with a confidence threshold of 65 percent.

To verify that the subspecies structure observed in the 3,888-loci ddRAD-seq dataset (fig. 3A) was retained in the targeted 93 nuclear SNPs based on Fluidigm genotype scores (validation goal 2), we completed a PCA implemented in the *adegenet* R package using the dudi.pca function and plotted the first two principal components. Additionally, we used STRUCTURE 2.2.3 to visualize genetic structure. The analysis was run for K=2 only, where K is the number of populations, as we were interested in determining the assignment probability as either a Tule or non-Tule goose. We also applied an admixture model with 100,000 burn-in and 1,000,000 MCMC iterations. The analysis was repeated five times. To assess the reliability of SNPtype assays for genotyping DNA from different DNA sources (tongue and feather extractions), the proportion of samples for each sample source yielding useable genotypic data was calculated in addition to any differences between sample sources (that is, allelic dropout, validation goal 3).

Results and Discussion

We obtained over 258 million raw sequencing reads with a maximum 150 bp length using single-end sequencing on two sequencing runs (three lanes in total) of an Illumina HiSeq 2500. The number of reads for each individual ranged from 344,091 to 1,643,653, with a median of 682,280. Initial exploration of genotyping results indicated that most loci were unambiguously genotyped across 238 of the 239 samples; one sample with a reduced proportion of high-quality genotypes was removed from the dataset. Additionally, preliminary PCAs led to the removal of four additional samples that seemed to diverge greatly from all others (data not shown). For the remaining 234 samples, a total of 4,155 polymorphic clusters (that is, putative single-copy loci) met the depth/genotype threshold. Of these loci, alignments for 2,939 passed automated checks for alignment quality, and an additional 949 loci passed thresholds after manual edits, for a total of 3,888 loci with good alignments that yielded 34,721 bi-allelic candidate SNPs or indels.

The PCA of these 3,888 loci indicated that the data were sufficient to genetically differentiate the Tule goose and those from other Pacific and Midcontinent Flyway GWFG populations (fig. 3A). This contrasts with an analysis of eight microsatellite loci (which suggested a panmictic population), although there were significant ($\alpha = 0.05$, adjusted for multiple comparisons using Benjamini and Yekutieli (2001) modified false discovery rate; see Table 2, Wilson and others, 2018) but small F_{ST} values between Tule goose and other subspecies (fig. 3B; see Wilson and others, 2018). The overall phi-st across all ddRAD-seq loci ranged from 0.025 to 0.078 between Tule goose and all other GWFG North American populations, depending on the pair of populations compared (table 2). Levels of divergence are similar to what others have found when comparing closely related species (Lavretsky and others, 2015).

Table 2. Lower triangular matrix estimating variance in allele frequencies between groups (phi-st values) calculated from 3,888 double-digest restriction-site-associated deoxyribonucleic acid sequencing loci for populations of greater white-fronted geese from different major geographic regions within the circumpolar distribution.

[Values within border indicate estimates restricted to North America. –s indicate that no comparison was done (no region was compared to itself). Blank cells are intentionally left blank because the comparisons are only done once]

Group	Eurasia	Greenland	Midcontinent	Pacific Flyway
Eurasia	–			
Greenland	0.051	–		
Midcontinent	0.005	0.046	–	
Pacific Flyway (non-Tule)	0.008	0.052	0.005	–
Cook Inlet Basin (Tule goose)	0.033	0.078	0.025	0.027

In North America, pairwise phi-st values were about five times greater for comparisons involving Tule goose than estimated between the Midcontinent and Pacific Flyway (BBL and YKD) GWFG populations (table 2), suggesting reduced genetic exchange between Tule goose and these other populations. Of the 3,888 loci, 165 (4.2 percent of total loci) and 144 (3.7 percent of total loci) showed elevated divergence (phi-st >0.1) between Tule goose and the Pacific Flyway (non-Tule) and Midcontinent GWFG populations, respectively (fig. 4). Only a limited number of loci (<1 percent) showed higher levels of divergence (phi-st > 0.2)—17 loci for Tule goose compared to non-Tule Pacific Flyway and 18 loci for Tule geese compared to Midcontinent. By comparison, only 5 loci had a phi-st value of 0.1 or greater when comparing non-Tule Pacific Flyway and Midcontinent GWFG populations. This shallow divergence suggests that the GWFG subspecies are more than likely recently diverged and (or) frequently intermix. Our ddRAD-seq analysis scanned only a small fraction of the genome; if subspecies diagnostic loci are present (that is, representing fixed differences between subspecies), a whole genomic approach likely would be required to detect them.

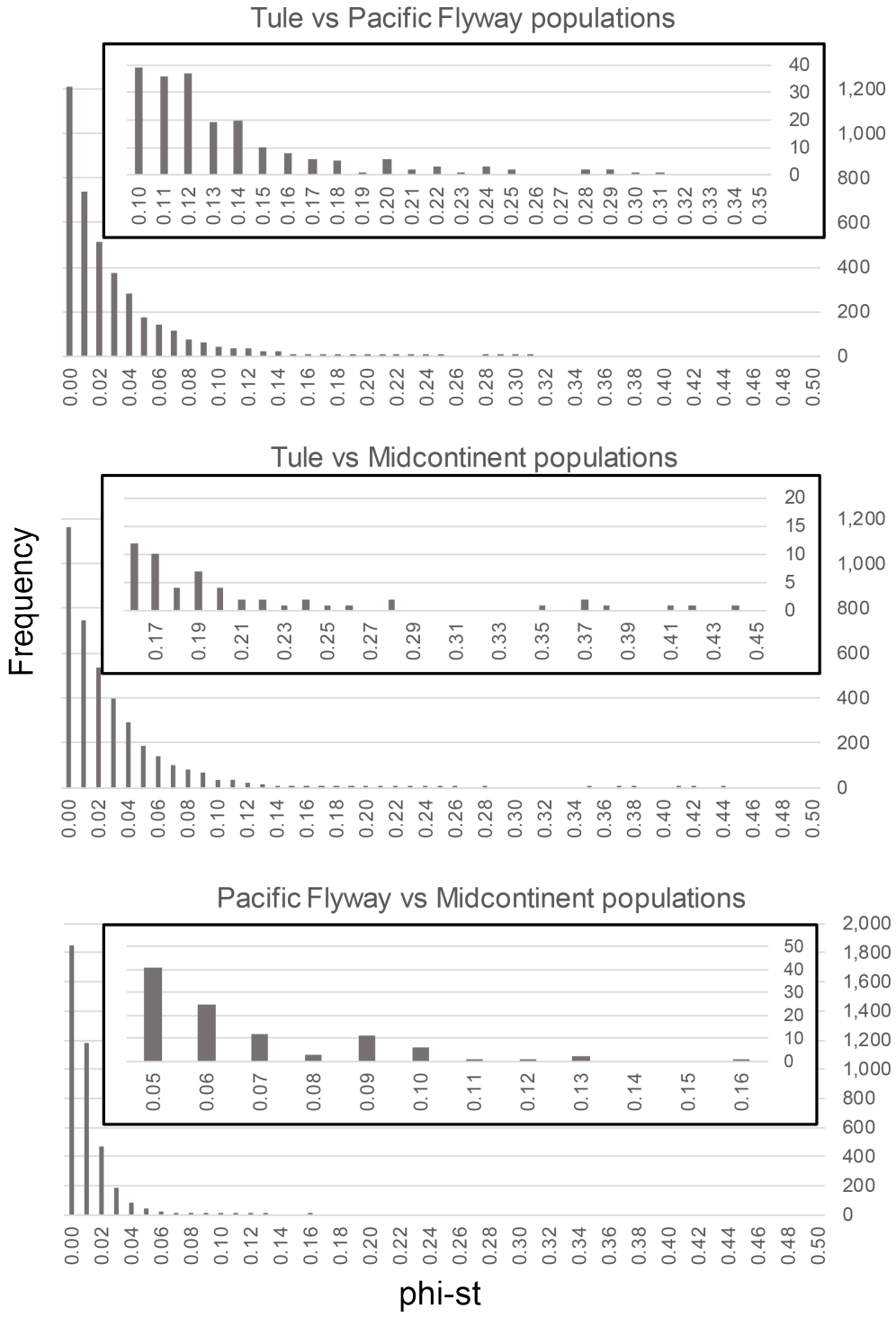

Figure 4. Graphs showing frequency (number of loci) of estimates of variance in allele frequencies between groups (phi-st) for 3,888 double-digest restriction site-associated deoxyribonucleic acid sequencing loci in comparisons of greater white-fronted geese (GWFG) North American populations. Pacific Flyway refers to non-Tule nesting populations of GWFG in Bristol Bay Lowlands and Yukon-Kuskokwim Delta, Alaska.

Validation of the Single Nucleotide Polymorphism Panel

PCA and STRUCTURE analysis of the 93 nuclear SNP loci (as genotyped from the ddRAD-seq data) using the reference dataset samples (see Wilson and others, 2019) are effective in discriminating between Tule geese and other GWFG North American populations (fig. 5), although there are some exceptions. As with the full dataset, one sample, B40 from CIB, putatively a Tule goose, was assigned with high probability as a non-Tule (blue cluster, deeper value). Morphological measurements of this goose indicated that this individual was not misidentified, as morphological measurements were within the range of other geese from CIB. Additionally, this sample had a mtDNA haplotype found in other Tule geese (Wilson and others, 2018). One Tule and non-Tule Pacific Flyway and two Midcontinent geese also had intermediate assignments (50–60 percent) in the STRUCTURE analysis, and therefore, could not be assigned to subspecies. Ambiguous placement of individuals that are morphologically identified as one subspecies but that are genetically more similar to another subspecies may be due to interbreeding.

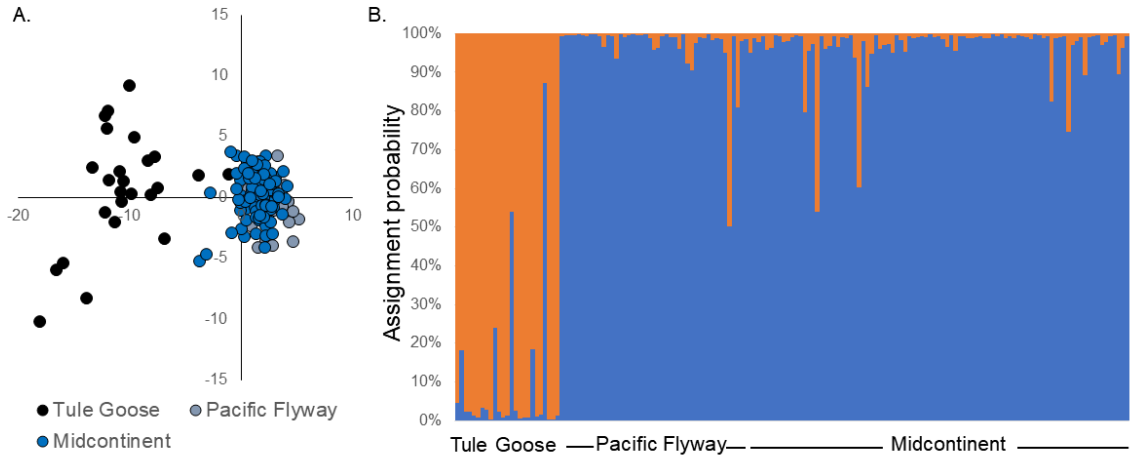

Figure 5. Scatterplot showing first two principal components (A), and graph showing average assignment probability of individual geese assigned to the two clusters by the program STRUCTURE (B) based on 93 nuclear single nucleotide polymorphisms (SNPs) genotyped using double-digest restriction site-associated deoxyribonucleic acid sequencing data and selected for inclusion in the 96-locus SNP panel for North American greater white-fronted geese (GWFG). Pacific Flyway refers to non-Tule nesting populations of GWFG in Bristol Bay Lowlands and Yukon-Kuskokwim Delta, Alaska. Numbers on x- and y-axes (A) are Principal Components Analysis coordinates and represent values associated with PC1 (x-axis) and PC2 (y-axis).

Verification of Genetic Structure of 93-Nuclear Fluidigm Single Nucleotide Polymorphism Panel

PCA and STRUCTURE analyses of the Fluidigm-derived genotypes showed a high level of genetic differentiation between Tule and non-Tule samples (fig. 6) similar to the reference ddRAD-seq dataset (fig. 5). As with the reference dataset (full and 93 SNP), a few samples were misassigned, potentially as a result of admixed ancestry or field misidentification. As in the ddRAD-seq data, B40 again was assigned as non-Tule, whereas a few other samples were inferred as being genetically admixed. In addition to B40, two other Tule geese (not included in ddRAD-seq dataset) also had genotypes characteristic of non-Tule Pacific Flyway genotypes

(fig. 6). This may indicate that there is immigration into the CIB from other nesting areas which may lead to the intermixing of subspecies as indicated by the intermediate assignment of some hunter-shot samples.

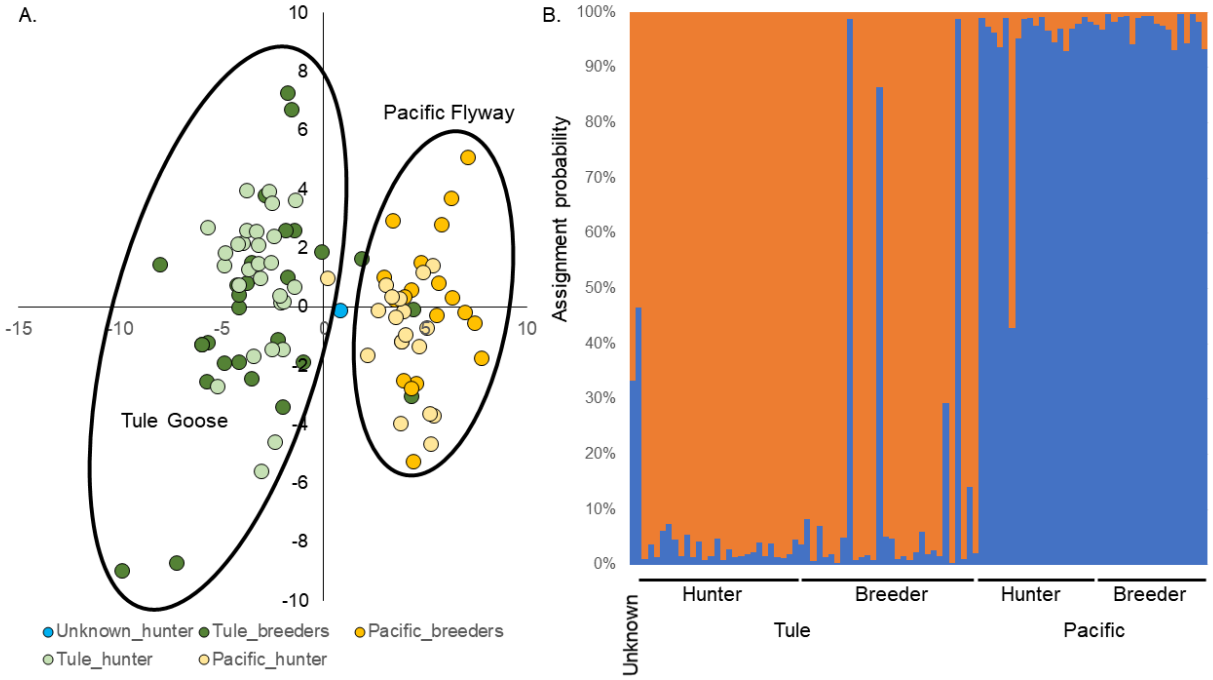

Figure 6. Scatterplot showing first two principal components (A) and graph showing STRUCTURE analysis (B) based on 93 nuclear single nucleotide polymorphisms using samples from known nesting areas (breeders) and California wintering areas (hunter). Hunter-shot samples were classified in the field to likely breeding areas based on phenotype. Pacific refers to non-Tule samples of greater white-fronted geese in the Pacific Flyway. Numbers on x- and y-axes (A) are Principal Components Analysis coordinates and represent values associated with PC1 (x-axis) and PC2 (y-axis).

Amplification Success of the Fluidigm Nuclear Single Nucleotide Polymorphism Panel

Run 1 (blood and tongue samples) yielded a median confidence genotype score of 99.99 (range 65.5–100) for successful samples across all nuclear SNPs (fig. 7). We also observed high amplification success—of a possible 8,835 genotypes (95 samples × 93 nuclear SNPs), we obtained 8,724 genotypes (98.7 percent). We also observed a high success rate within each targeted SNP, with the number of failed samples (genotype scores less than 65 percent) ranging from 0 to 8 out of a total of 95. The SNP with eight failed samples contained an indel (G/-). Amplification success also was greatly increased using the pre-amplification step (STA assay), as most blood samples in Run 2 failed without this step but amplified in Run 1. Feather samples in Run 2 did not perform as well as the blood or tongue samples from Run 1 across all SNP assays. For feathers overall, 45 SNP assays (40 percent) failed completely (that is, all samples failed) and in 41 SNP assays (44 percent), at least one-half of the samples failed (fig. 7).

We compared SNP genotypes obtained from DNA extracted from feathers and tongue, both sampled by hunter-check station personnel from the same individual geese harvested by hunters on wintering grounds in California, to examine the reliability of SNP genotypes observed from different tissue sources. Because the quantity of DNA extracted from feathers usually is less than the quantity extracted from blood or muscle tissue, we tested for low reliability (for

example, genotyping errors) and poor performance (amplification success) of each sample type. In addition to lower amplification success for feather samples across all SNP assays, we observed an error rate (recovery of different genotypes between different sample types from the same individual) of 12 percent (120 differences out of 985 total genotype calls), based on nuclear SNPs. Of these differences, 87.5 percent could be attributed to allelic dropout in feathers (that is, the feather genotype was scored as homozygous, whereas the tongue genotype was scored as heterozygous). Allelic dropout is a common occurrence in SNP amplification from low-quantity and (or) low-quality DNA samples (Bayerl and others, 2018). Average DNA concentration for the feather samples was 29.9 ng/µL (median 28.1 ng/µL), whereas blood extracts averaged 511.2 ng/µL (median 409.5 ng/µL) and tongue extracts averaged 818.3 ng/µL (median 836.4 ng/µL). Blood and tongue extractions were diluted to 50 ng/µL working solutions as recommended by Fluidigm except for two blood samples that had concentrations of 1.2 ng/µL and 20.3 ng/µL. Those lower concentration samples successfully amplified 91 and 92 of the nuclear SNP assays, respectively; therefore, it seems that the lower amplification success of feathers is due to low-quality DNA.

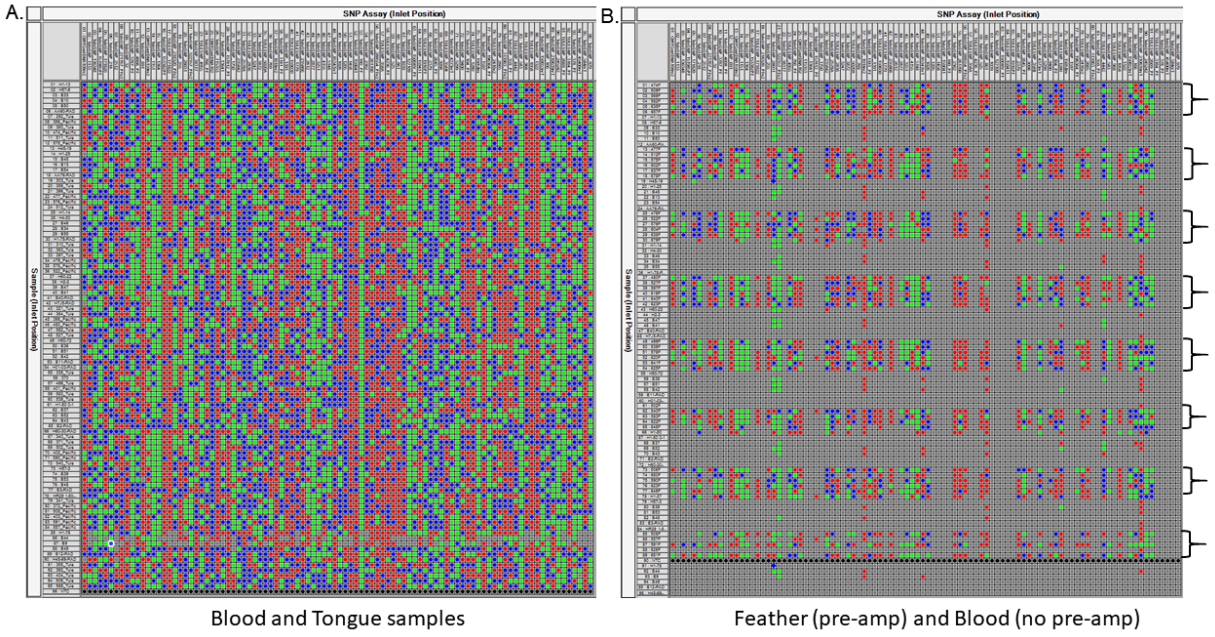

Figure 7. Image representing amplification success of single nucleotide polymorphism assays using (A) blood and tongue samples with pre-amplification step, and (B) feather with pre-amplification step and blood without pre-amplification. Brackets on right panel indicate which samples are feathers. Green, blue, and red pixels (darker values) indicate a successful amplification. Gray pixels (light values) indicate failed sample, with black pixels indicating positive control sample.

Amplification Success of the Fluidigm Mitochondrial Deoxyribonucleic Acid Single Nucleotide Polymorphism Panel

In addition to testing for amplification success of assays in tongue or blood and feather samples, we also tested for the possibility that SNP primers were amplifying nuclear pseudogenes of mitochondrial origin. Because mitochondrial DNA is single-stranded and nuclear DNA is double-stranded, any evidence of heterozygous SNPs would indicate pseudogenes.

Amplification of nuclear pseudogenes rather than mitochondrial DNA would result in erroneous genotypes for that marker and may lead to incorrect assignments of individuals to subspecies. In Run 1, we observed a percentage of samples called as heterozygotes in the three mtDNA SNPs—(1) 24 percent in SNP locus mtDNA106rv2, (2) 18 percent in SNP locus mtDNA159rv2, and (3) 84 percent in SNP locus mtDNA162. For SNP mtDNA106rv2 and mtDNA159rv2, all heterozygote calls were from blood samples, whereas in mtDNA162, both blood and tongue accounted for a high percentage of heterozygotes. In contrast, there was only one heterozygote genotype call in mtDNA159rv2 for the feather samples (fig. 8).

Nuclear pseudogenes of mitochondrial origin are a common occurrence in avian species (Lopez and others, 1994; Sorenson and Quinn, 1998) and the observation of heterozygotes for mtDNA SNPs is most likely due to the presence of both a mitochondrial SNP and a nuclear pseudogene allele. This hypothesis is supported by the observation of a higher rate of heterozygous genotypes in blood samples relative to feather and tongue samples. Avian red blood cells are nucleated, and when mature lack mtDNA such that avian blood samples yield a relatively small amount of mitochondrial DNA relative to nuclear DNA. In contrast, almost all cells in feather and tongue samples have multiple mitochondria per cell, yielding a higher ratio of mtDNA to nuclear DNA. As such, we recommend that if the three mitochondrial DNA SNPs are included in future panels, DNA from feather samples be used to generate those SNPs. Additionally, if tongue samples are used SNP mtDNA162 should be excluded from that analysis and any heterozygous genotype calls for other mtDNA SNPs should be excluded.

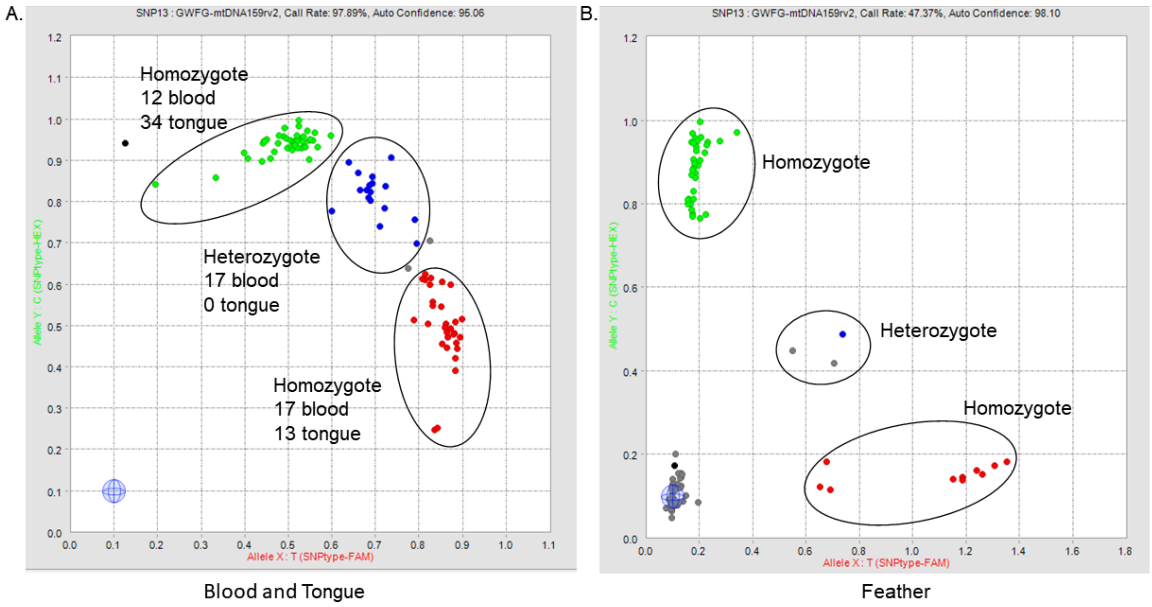

Figure 8. Scatterplots showing results of testing for presence of pseudogenes in greater white-fronted goose mitochondrial deoxyribonucleic acid (mtDNA) single nucleotide polymorphisms (SNPs) from different sample sources—blood and tongue (A), and feather (B). Scatterplots show genotype calls of SNP mtDNA159rv2. This scatterplot shows the high frequency of heterozygote genotype calls when using blood samples as compared to feather samples.

Summary

We showed that the 93-locus nuclear single nucleotide polymorphism (SNP) panel was effective in discriminating Tule geese from individuals of other Pacific Flyway greater white-fronted geese populations in about 95 percent of cases, assuming original assignments based on geography and (or) phenotype were correct. The cases of misclassification based on SNP analysis could be due to initial misidentification at hunter-check stations, and (or) intermixing of the two Pacific Flyway subspecies. Addition of the three mtDNA SNPs may add resolving power to the panel, but care should be taken to avoid the inclusion of heterozygous genotypes in data analyses, as they likely represent a combination of a true mtDNA allele and a nuclear pseudogene. The 96-locus SNP assay panel designed for the Fluidigm Corporation Genotyping System can be used for in-season harvest identification during future hunting seasons. We note that SNPs identified and used in panels developed for one platform (here, the Fluidigm) can be easily adapted for use in other systems.

Data Availability

All data supporting the findings in this report are available in a USGS data release (Wilson and others, 2019).

Acknowledgments

We thank Melanie Weaver and Dan Skalos (State of California, Department of Fish and Wildlife) for their assistance in gathering hunter-harvested samples and their input in the original design of this study, Kevin Sage (U.S. Geological Survey [USGS]) for laboratory guidance, and Barbara Pierson (USGS) for assistance in data archiving. We are thankful for the tremendous support of colleagues from around the world who collected samples for us at various locations.

References Cited

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., 1990, Basic local alignment search tool: Journal of Molecular Biology, v. 215, p. 403–410.

Bay, R.A., Harrigan, R.J., Underwood, V.L., Gibbs, H.L., Smith, T.B., and Ruegg, K., 2018, Genomic signals of selection predict climate-driven population declines in a migratory bird Science, v. 359, p. 83–86.

Bayerl, H. Kraus, R.H.S., Nowak, C., Foster, D.W., Fickel, J., and Kuehn R., 2018, Fast and cost-effective single nucleotide polymorphism (SNP) detection in the absence of a reference genome using semideep next-generation Random Amplicon Sequencing (RAMseq): Molecular Ecology Resources, v. 18, p. 107–117.

Benjamini, Y., and Yekutieli, D, 2001, The control of false discovery rate in multiple testing under dependency: Annals of Statistics, vol. 29, no. 4, p. 1165–1188.

Contina, A., Bay, R.A., Underwood, V.L., Smith, T.B., Kelly, J.F., Bridge, E.S., and Ruegg, K.C., 2017, Characterization of SNP markers for the painted bunting (*Passerina ciris*) and their relevance in population differentiation and genome evolution studies: Conservation Genetics Resources, v. 11, no. 1, p. 5–10.

DaCosta, J.M., and Sorenson, M.D., 2014, Amplification biases and consistent recovery of loci in a double-digest RAD-seq protocol: PLoS ONE, v. 9, no. 9, p. e106713.

Deuel, B., and Takekawa, J.Y., 2008, Studies of western birds: Western Field Ornithologist's Union, Camarillo, California, and California Department of Fish and Game, Sacramento, California.
Dray, S. and Dufour, A-B., 2007, The ade4 package—Implementing the duality diagram for ecologists: Journal of Statistical Software, v. 22, p. 1–20.
Edgar, R.C., 2004, MUSCLE—Multiple sequence alignment with high accuracy and high throughput: Nucleic Acids Research, v. 32, no. 5, p. 1792–1797.
Edgar, R.C., 2010, Search and clustering orders of magnitude faster that BLAST: Bioinformatics, v. 26, no. 19, p. 2460–2461.
Ely, C.R., 2008, Maintenance of population structuring in sympatric-wintering populations of greater white-fronted geese *Anser albifrons*—Behavior, ecology, and landscapes: Vogelwelt, v. 129, p. 310–316.
Ely, C.R., and Dzubin, A.X., 1994, Greater white-fronted goose (*Anser albifrons*), *in* Poole, A., and Gill, F., eds., The birds of North America, No. 131: The Academy of Natural Sciences, Philadelphia, and The American Ornithologists' Union, Washington, D.C.
Ely, C.R., Fox, A.D., Alisaukas, R T., Andreev, A., Bromley, R.G., Degtyarev, A.G., Ebbinge, B., Gurtovaya, E.N., Kerbes, R., Kondratyev, A.V., Kostin, I., Krechmar, A.V., Litvin, K., Miyabayashi, Y., Mooij, J.H., Oates, R.M., Orthmeyer D.L., Sabano, Y., Simpson, S.G., Solovieva, D.V., Spindler, M.A., Syroechkovsky, Y.V., Takekawa, J.Y., and Walsh, A., 2005, Circumpolar variation in morphological characteristics of greater white-fronted geese (*Anser albifrons*): Bird Study, v. 52, p. 104–119.
Ely, C.R., and Takekawa, J.Y., 1996, Geographic variation in migratory behavior of greater white-fronted geese (*Anser albifrons*): The Auk, v. 113, p. 889–901.
Ely, C.R., Wilson, R.E., and Talbot, S.L., 2017, Genetic structure among greater white-fronted goose populations of the Pacific Flyway: Ecology and Evolution, v. 7, p. 2956–2968.
Jombart, T., 2008. *adegenet*—A R package for the multivariate analysis of genetic markers: Bioinformatics, v. 24, p. 1403–1405.
Lavretsky, P., DaCosta, J.M., Hernandez-Baños, B.E., Engilis, A., Jr., Sorenson, M.D., and Peters J.L., 2015, Speciation genomics and a role for the Z chromosome in the early stages of divergence between Mexican ducks and mallards: Molecular Ecology, v. 24, p. 5364–5378.
Lopez, J.V., Yuhki, N., Masuda, R., Modi, W., and O'Brien, S.J., 1994, *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat: Journal of Molecular Evolution, v.3, p. 174–190.
Mensik, G. 1991. Pacific Flyway management plan for the Tule greater white-fronted goose: Pacific Flyway Study committee, U.S. Fish and Wildlife Service report, Portland, Oregon.
Novembre, J., and Stephens, M., 2008, Interpreting principal component analyses of spatial population genetic variation: Nature Genetics, v. 40, p. 646–649.
Olson, S.M., compiler, 2014, Pacific Flyway data book—Migratory bird population indices, harves, and hunter participation and success: U. S. Fish and Wildlife Service, Division of Migratory Bird Management, Vancouver, Washington. [Also available at http://pacificflyway.gov/Documents/Databook.pdf.]
Orthmeyer, D.L., Takekawa, J.Y., Ely, C.R., Wege, M.L., and Newton, W.E., 1995, Morphological differences in Pacific Coast populations of greater white-fronted geese (*A. albifrons*): Condor, v. 97, p. 123–132.
Pritchard, J.K., Stephens, M., and Donnelly, P., 2000, Inference of population structure using multilocus genotype data: Genetics, v.155, p. 945–959.

Sorenson, M.D. and Quinn, T.W. 1998, Numts—A challenge for avian systematics and population biology. The Auk, v. 115, p. 214–221.
U.S. Fish and Wildlife Service, 2014, Migratory bird hunting regulations on certain federal Indian reservations and ceded lands for the 2014-15 early season; Federal Register, V. 79, P. 52226–52238.
Wilson, R.E. Ely, C.R., and Talbot. S.L., 2018, Flyway structure in the circumpolar greater white-fronted goose: Ecology and Evolution, v. 8, p. 8490–8507.
Wilson, R.E., Pierson, B.J., Sonsthagen, S.A., Ely, C.R., and Talbot, S.L., 2019, Development of single nucleotide polymorphisms (SNPs) in greater white-fronted geese (*Anser albifrons*) for genetic stock identification on wintering grounds, 2019: U.S. Geological Survey data release, https://doi.org/10.5066/P9LYUFRH.

Publishing support provided by the U.S. Geological Survey
Science Publishing Network, Tacoma Publishing Service Center

For more information concerning the research in this report, contact the
Director, Alaska Science Center
U.S. Geological Survey
4210 University Drive
Anchorage, Alaska 99508
https://www.usgs.gov/centers/asc/

ISSN 2331-1258 (online)
https://doi.org/10.3133/ofr20191040