A PHYLOGENOMIC ANALYSIS OF THE TRUMPETER (CYGNUS BUCCINATOR) AND TUNDRA SWAN (CYGNUS COLUMBIANUS COLUMBIANUS)

by

Christopher K. Yesmont

A Thesis
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Master of Science
Biology

Committee:

_____	Dr. Patrick M. Gillevet, Thesis Director
_____	Dr. Ancha Baranova, Committee Chair
_____	Dr. Dmitri Klimov, Committee Member
_____	Dr. James D. Willett, Director, School of Systems Biology
_____	Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science

Date: 7/28/14

Summer Semester 2014
George Mason University
Fairfax, VA

A Phylogenomic Analysis Of The Trumpeter (*Cygnus buccinator*) And Tundra Swans (*Cygnus columbianus columbianus*)

A Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at George Mason University

by

Christopher Kenneth Yesmont
Bachelor of Science
Stony Brook University, 2007

Director: Patrick M. Gillevet, Professor
Department of Biology

Summer Semester 2014
George Mason University
Fairfax, VA

DEDICATION

I dedicate this Thesis to my family, and especially to Kathleen Grimley for sticking by my side during the ups and downs of my education.

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Baranova for guiding me through the logistics of my graduate education and for giving me the opportunity to work with Dr. Gillevet on the The Swan Project. I especially thank Dr. Pat Gillevet, my Thesis Director, for being tremendously helpful and optimistic throughout this research. Thank you to Dr. Masoumeh Sikaroodi, for without her this research would not have been possible. Much appreciation for all of those who worked on The Swan Project before me, Lauren Wilson, M.S., and Elizabeth Dingess, M.S. Finally, I cannot begin to express my gratitude to my friends and family for being so understanding and supportive during this intense period of study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

A PHYLOGENOMIC ANALYSIS OF THE TRUMPETER (CYGNUS BUCCINATOR) AND TUNDRA SWANS (CYGNUS COLUMBIANUS COLUMBIANUS)

Christopher K. Yesmont, M.S.

George Mason University, 2014

Thesis Director: Dr. Patrick M. Gillevet

High-throughput or Next-Generation sequencing (NGS) has proven to be a time-efficient and cost-effective technique for the sequencing of whole-genomes. Despite its benefits, high-throughput sequencing is remaining a relatively nascent technique for phylogenetic analyses (McCormack 2013). We performed sequence assembly and analysis of genomic reads from a wild Tundra Swan (*C. columbianus*) and a wild Trumpeter Swan (*Cygnus buccinator*). The reads were assembled into Contigs that were later run through a custom Perl script designed to extract the differences between sequences. We also performed further analysis of the genetic relatedness between swan species using sequences of nuclear gene encoding the chromo-helicase-DNA binding domain (CHD) and associated introns that involved alignments and phylogenetic tree construction. Additionally, polymorphisms in the mitochondrial control region (D-loop) of Trumpeters, Tundra Swans, and hybrids were used for phylogenetic trees analysis. The

phylogenetic analysis of the nuclear gene (*CHD*), D-loop, and genomic sequences indicates strong introgression between the two species due to extensive hybridization.

INTRODUCTION

The Trumpeter Swan (*C. buccinator*) and Tundra Swan (*Cygnus columbianus columbianus*) are two migratory large waterfowl species native to North America. Along with Bewick's swan (*C. bewickii*), and whooper swan (*C. Cygnus*), these species form an evolutionary complex known as the four northern swans of the family *Anatidae*. The Trumpeter and the Tundra species are further grouped together, but the relationships among the northern swans within the family have not been resolved (Travsky and Beauvais 2004).

The Trumpeter swan was abundantly distributed throughout North America prior to the 20th century, but overhunting and harvesting of the swans for their plumage reduced the species to numbers near extinction (Travsky and Beauvais 2004). In 1935, the establishment of protected breeding grounds such as Red Rock Lakes and a moratorium on hunting led to a rebound in the Trumpeter swans' number. Currently, the Trumpeter swans are divided into three populations that reflect migration corridors of the species; Pacific Coast Population (Alaska, British Columbia, Yukon Territory), the Rocky Mountain Population (shared boundaries of Idaho, Montana, and Wyoming), and the Interior Population (east of the Rocky Mountains) (Barret and Vyse 1982, Oyler-McCance et al. 2007). It

is important to note that these population designations do not define barriers to gene flow in the populations (Oyler-McCance et al. 2007). Recent mitochondrial DNA analysis of the Rocky Mountain Population and the Pacific Coast Population by Oyler-McCance et al. (2007) demonstrated decreased genetic diversity compared to other waterfowl species consistent with a historic bottleneck of the Trumpeter species.

Historically, the breeding ranges of these two swan species were isolated from each other, with the Trumpeter's breeding range restricted to the boreal forest (taiga) habitat, while the Tundra's range covers the eastern Canadian Arctic to Bristol Bay, Alaska (Wilk 1988). Changes in the sub-arctic vegetative composition, anthropogenic disturbances, and increases in annual temperatures due to climate change have led to an overlap in breeding grounds of the two species (Beck et al. 2011, Schmidt et al. 2011, Schmidt et al. 2009).

Importantly, Trumpeter and Tundra swans have the ability to produce fertile offspring through interbreeding. This hybridization has been documented in captivity, and a few cases have been documented in the wild (Travsky and Beauvais 2004). Interbreeding is of conservation concern because genetic mixing resulting from hybridization can lead to the extinction of a species already in decline when that species becomes "genetically swamped" by such introgression (Dabrowski et al. 2005). If the Trumpeter and Tundra swans are actively interbreeding, the overlapping of breeding ranges brought on by anthropogenic effects could pose serious threats to the conservation of these two species.

This potential hybridization in the wild necessitates the development of techniques allowing distinguishing the Tundra's from the Trumpeters as well as each of the species from their hybrids. Some recent works have concentrated on this aspect of swan conservation (St. John *et al.* 2006, Wilson 2013).

With the advent of Next-generation sequencing (NGS) technology, the field of phlyogenomics, or the comparative analysis of genome-scale data that aims to infer evolutionary relationships between species and populations (Chan and Ragan 2013), has flourished. High-throughput sequencing is a time-efficient and cost-effective technique for the sequencing of the genomes of endangered species. Despite the obvious benefits of NGS, its use in phylogenetic studies relevant to species ecology and conservation biology has been limited, and somewhat slow to catch on in comparison to high demand areas such as metagenomics and clinical genetics (McCormack 2013).

The objective of this study is to compare two swan species, Tundra (*C. columbianus*) and Trumpeter (*Cygnus buccinators*), from a phylogenomics perspective. We aim to look for the overall sequence differences and to reveal just how similar or different the species really are. Next-Generation or high throughput sequencing is a suitable technique of choice, because it allows for whole-genome comparisons that are not anchored to prior knowledge of the genomes being sequenced.

Taxonomy

Present consensus places Trumpeter (*Cygnus buccinators*), and Tundra (*C. columbianus columbianus*) swans in the tribe *Anserini* that consists of three genera of the *Anatidae* family, otherwise known as the waterfowl (ducks, swans, geese). *Anserini* tribe comprises the largest waterfowl species (Baldassarre and Bolen, 1994). Trumpeter and tundra swans belong to an evolutionary complex commonly referred to as the Northern Swans. This name was suggested by Delacour (1954) in order to distinguish the "knob-less" white swan species breeding in the northern hemisphere from similar swan species in the southern hemisphere. Included in this group are the Eurasian counterparts to trumpeter and tundra, whooper swan (*C. Cygnus*) and Bewick's swan (*C. bewickii*).

The evolutionary relationships between the Northern Swan species are not yet established. However, attempts to resolve the relationship have demonstrated a close-knit relationship between swans in the grouping. For instance, through DNA fingerprinting Meng and Perkin (1993) showed a high degree of relatedness between trumpeter, whooper, and Bewick's swans. The high level of molecular, along with morphological similarity has led to varying proposals regarding taxonomic placement of these swans. Banko and Schorger (1976) have suggested the existence of a superspecies made up of Whooper and Trumpeter swans while others have regarded the Trumpeter as a subspecies of the Whooper swan. Others such as Portenko (1972) and Johnsgard (1974) considered Trumpeter and Tundra to be conspecific.

Description

All Northern Swans exhibit similarity in their exterior traits, including mostly white downy plumage, large size, strong vocalizations, and a lack of sexual dimorphism and dichromatism (Baldassarre and Bolen 1994; Delacour 1954). Bill coloration, specifically the amount of yellow in the bill, is a major distinguishing feature for identification of particular type of Northern swans in the field (Evans and Sladen 1980). The Eurasian species exhibit significantly more yellow on their bills than their New World counterparts (Evans and Sladen 1980), while the trumpeter swans have bills that are completely black in color (Banko 1960). The Tundra' bills are predominantly black with a small patch of yellow in the front of the eye (Evans and Sladen 1980). Pink-billed and yellow-lored variants of Trumpeter swans have been documented in Yellowstone National Park.

The Trumpeter has a wingspan ranging between 2.1 and 2.4 m and weighs between 10 and 15 kg. These measurements make it the largest waterfowl in the world. Adult trumpeters have an entirely white downy plumage, black bill, and black feet and tarsi. Often their head and neck exhibit a rusty stain as a result of feeding in ferrous waters. Cygnets (swan' juveniles) have a brownish-gray growth phase. Similar to adults, they sometimes possess rust-stained plumage. Contrary to adults, however, their feet are yellowish or olive gray-black in color. Cygnet bills are mostly black with the exception of the basal portion of the culmenary

ridge behind the nostril that is a light pink in its color. Newly hatched cygnets are completely white, with yellow feet and pink bills (Travsky and Beauvais 2004).

A major distinction between Trumpeters and Tundra swans is in their internal anatomical morphology. Trumpeters have a convoluted trachea inside the sternum with a mid-dorsal protrusion (Johnsgard 1978). This extra loop in the trachea gives the trumpeter a much deeper voice allowing it's call to be distinguished from that of the tundra swan in the wild (Banko 1960). Frequent vocalizations are common in the Trumpeter. However, during nesting and brooding seasons pairs of Trumpeters are relatively silent. Vocalization resumes in the fall and is both individually expressed and synchronized in groups until March and April when they reach a climax (Banko 1960).

Habitat

Trumpeters are known to occupy a wide variety of habitats. However, both the current observations and historical observations from fur-trade records have led to the consensus that the preferred habitat of the trumpeter swan is the open boreal forest (Banko 1960). Because of the reproductive cycle requirement of 140-154 ice-free days, there is a limit of 2,700 ft. on the elevation of suitable habitat (Travsky and Beauvais 2004). Additional breeding habitat requirements are stable waters that lack seasonal fluctuations, for example lakes, marshes, and sloughs. It is important that the water is shallow enough to permit foraging on aquatic plants such as crowfoot (*Ranunclulu*), and tubers of arrowhead

(*Sagittaria*) (Johnsgard 1978). Nests are built on vegetation or small islands and Muskrat and beaver lodges are sometimes used as nest substrate (Travsky and Beauvais 2004). In a survey of water basins used for nesting by Trumpeter swans in Alaska, Hansen et al. (1971) found that 51 percent of nests were in beaver impoundments.

As one might expect, wintering habitats of the trumpeter are distinct from their breeding habitats. Open water adjacent to level and open terrain is a key requirement for wintering sites. Open terrain allows for adequate space for takeoff and landing of such a large waterfowl. It also prevents impairment of vision when Trumpeters are resting for the winter.

In contrast to the boreal habitat of the trumpeter swan, as their name would imply, the Tundra swan inhabits the open tundra. Another difference in habitat preference between trumpeter and tundra swans is that tundra swans prefer upland nest sites and does not rely on emergent vegetation for nest building (Hansen et al. 1971). Hansen et al. (1971) described the tundra swan nests as resembling a “volcanic cone”, while the Trumpeter nests were described as resembling a “bulky, round haystack”. These distinct features have allowed surveyors to differentiate between nest sites of the two species in regions where breeding grounds overlap.

Tundra swans feed on aquatic plants such as wild celery (*Vallisneria),* wigeon grass, bulrushes, pondweeds, and *Sagittaria* (Johnsgard 1978). In brackish

waters, tundra swans have been known to feed on mollusks such as clams (Johnsgard 1978).

Geographic distribution

Trumpeter Swan

The Trumpeter had a vast prehistoric distribution across the North American continent. Trumpeter remains have been identified in widely separated geologic formations dating back to the Pleistocene in regions such as; Aurora Illinois; Itchtucknee River, Florida; and Fossil Lake, Oregon (Banko 1960). Trumpeter remains have been found alongside bones of the giant beaver and mastodon in Illinois, and the Californian condor (*Gymnogyps californianus*), whooping crane (*Grus americana*), and jabiru stork (*Jabiru mycteria*) in Florida. Such findings in habitats considered unsuitable for modern-day Trumpeters represent not only changes in climate, geology and ecology, but also the ability of ancient trumpeter-like swan species to adapt to a changing environment.

Similar to the extensive prehistoric range of the Trumpeter swan, the historic range is believed to be much greater than the current distribution since Trumpeters were extirpated from most of their historic range (Banko 1960; Shea et al. 1991). However, brief and scattered accounts in the literature over a long period of time make it difficult to know the exact distribution of the historic range (Banko 1960). The first account to separate the Trumpeter as a distinct species of

swan and report it's occurrence on the east coast of the US was made by John Lawson, a Surveyor-General of North Carolina (Banko 1960).

Hunting and exploitation of the Trumpeter swan peaked in the late 1800's and caused a major decline in populations that would reduce the species to numbers near extinction (Banko 1960). Some of the earliest evidence of the harvesting of Trumpeters comes from bibliographical literature by E. S. Thomas, Curator of Natural History of the Ohio Historical Society in which he recounts unearthing bones of the swan among kitchen material from Native American archeological sites. However, it seems that only with the arrival of European settlers colonizing North America, and the realization of the value of swan plumage, quills, and eggs in commerce with European markets did the real decimation of Trumpeter populations begin. Swan plumage was apparently used in the manufacturing of powder puffs (Delacour 1954), eggs valued by collectors, and quills supposedly made some of the best pens, especially for illustrators. Ironically, John James Audobon preferred Trumpeter quills for illustrating small birds (Banko 1960).

Commercialization of swan hunting took shape when large companies such as Hudson's Bay Co. and Canadian Co. began exporting large quantities of plumage, quills and wings to markets in London (Banko 1960). These two companies later merged under the name Hudson's Bay Company. According to historical documents, Hudson's Bay Company exported a total of 17,671 skins between the years 1853 and 1877. The average annual export of skins was at its

maximum from 1853 to 1867, but this was followed by a sharp decline in the last seven years between 1870 and 1877 (Banko 1960). This decline, coupled with the relatively high price for trumpeter eggs in the last decade of the 19th century implicates Hudson's Bay Company and their practices as the major contributor in the extirpation of the species.

The Migratory Bird Treaty Act of 1918 made hunting of Trumpeters illegal, except by this time the species was thought to be extinct by ornithologists and authorities of the like. In 1932 surveys of the species found only 69 individuals in the Tristate region of Montana, Idaho and Wyoming (Shea et al. 1991). Attempts to reintroduce into historic ranges by establishment of refuges such as the Red Rock Lakes National Wildlife Refuge, Montana in 1935 saw rebounds in populations (Banko 1960; Oyler -McCance et al. 2007). Shortly after, several thousand pairs of Trumpeters were discovered in Alaska in 1954 (Travsky and Beauvais 2004).

The current distribution of Trumpeter is essentially three separate populations. These populations are the Pacific Coast Population (PCP), the Rocky Mountain populations (RMP) and the Interior Population (IP). The Pacific Coast Population includes birds nesting in Alaska and wintering along the coast of Canada and Pacific Northwest United States. The RMP is made up of the shared boundaries of Idaho, Montana, and Wyoming. The RMP is further divided into three discrete flocks of Trumpeters based on wintering location. These flocks are the Canada flock, tri-state flock and the restoration flock (Barret and Vyse 1982).

The restoration flock is the result of translocations of adults and cygnets from the tri-state flock to refuges. The Interior Population contains birds that nest in the range east of the Rocky Mountains and is also the result of restoration efforts (Barret and Vyse 1982, Oyler-McCance et al. 2007).

Due to the near decimation of the species followed by the population rebounds stemming from such low numbers, Trumpeters underwent a significant bottleneck in the early 20th century and exhibit low genetic diversity throughout their range as a result (Barrett and Vyse 1982; Meng et al. 1990; Marsolais and White 1997; Oyler-McCance et al. 2007). When examining mitochondrial and nuclear markers, Oyler-McCance et al. (2007), found a significant differentiation between the Pacific Coast populations and the eastern RMP populations. They also found there to be more genetic structure in the Pacific coast populations. Interestingly, their mitochondrial DNA analysis revealed a low degree of variability compared to other waterfowl suggesting a bottleneck previous to the documented bottleneck led by the reintroduction of the species.

Tundra Swan

The Tundra Swan (*Cygnus columbianus columbianus*) is one of the most common swan species of North America. For management purposes, Tundra's are divided into two populations based on their wintering distribution; an East Coast population (EP) and a West Coast population (WP). The Tundra swan was formerly referred to as Whistling swan because of the swan's characteristic

whistling and is further differentiated by its “bark-like” vocalizations (Johnsgard 1978). U.S. Fish and Wildlife Service estimates from 1989 found a total of 169,300 Tundra Swans wintering in the US (Limpert et al. 1991). As a result of the high population abundance relative to other swans, the ban on hunting Tundra Swans in the Western population was lifted in 1962 and subsequently lifted from the Eastern population in 1984 (Serie and Bartonek 1991). This estimate is an increase from the prior surveys and the species is believed to be gradually increasing in number. Contrary to the Eastern population, the Western population seems to have declined slightly through hunting, habitat degradation and poisoning as a result of lead shot ingestion. The breeding range of *C. columbianus* extends from Hudson Bay in Canada westward across the arctic tundra to the Bering Sea (Limpert et al. 1991). The primary nesting area in Alaska is on the Yukon-Kuskokwim delta and the major nesting areas in Canada are the Mackenzie and Anderson River deltas of Canada (Limpert et al. 1991). In 19th century, extensive hunting caused a reduction in populations of tundra’s, but this decline was not as drastic as seen in the Trumpeter swan.

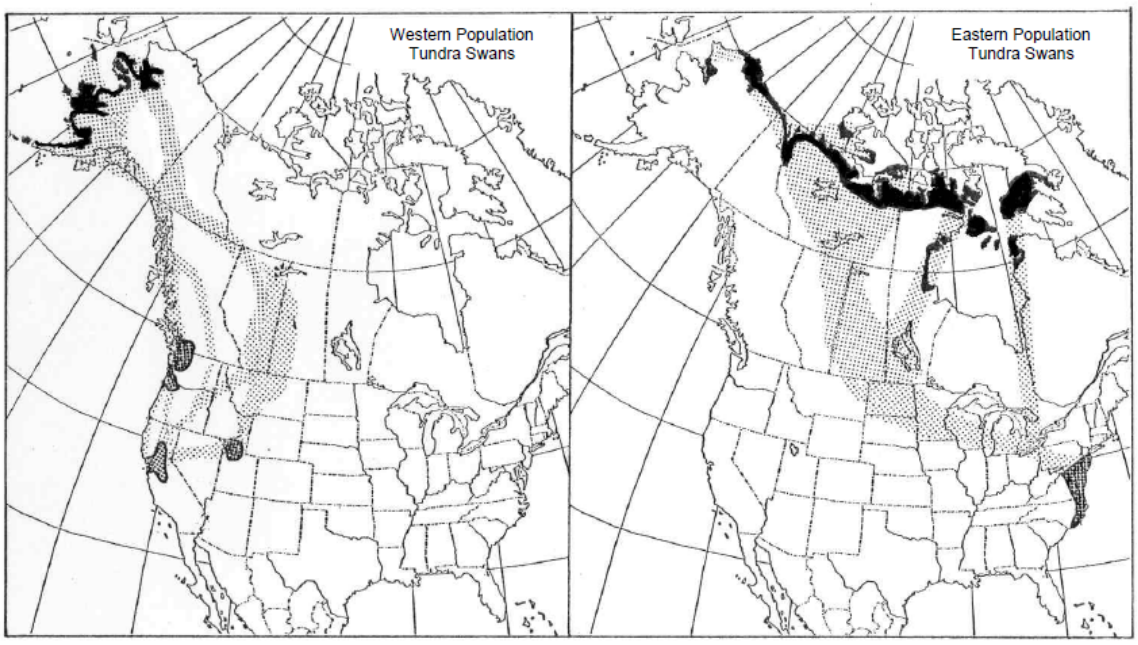

Figure 1. The Distribution of Eastern and Western Populations of Tundra Swans. According to FWS 2001.

Hybridization of Swan Species

Hybridization, or the interbreeding of species, is common among avian species. As much as 10% of bird species are known to hybridize in the wild (Grant and Grant 1992). Some of the typical avian hybrid relationships have been discovered in the passerines (Dabrowski et al. 2005; Faivre et al. 1999; Stein and Uy 2006), seabirds (Pons et al. 2013) and many others (McCarthy 2006). In swans, hybridization in captivity outnumbers the instances of hybridization in the wild. However, several cases of swan species interbreeding in the wild have been

documented in *Cygnus cygnus* (whooper) with *Cygnus olor* (mute swan) (Panov and Pavlova 2010; McCarthy 2006), *Cygnus atratus* (Black Swan) with *Cygnus olor* in Australia and New Zealand, *Cygnus bewickii* with *Cygnus columbianus* (Tundra), and finally *Cygnus buccinators* (Trumpeter) with *Cygnus columbianus* (McCarthy 2006).

In captivity, hybridization between Trumpeter and Tundra swans result in fertile offspring. These hybrids can be found at Airlie Swan Research center, Warrenton, Virginia. The F1 generation offspring of the cross between Trumpeter and Tundra are referred to as Trumplings, whose name is a derivation of **Trump**eter and Whist**ling** swan (Sladen, 2007). Backcrosses of Trumpeters with Trumplings occur as well, and are referred to as Trumpetlings.

Hybridization is of concern to the conservation of species because extinction of one species may occur through the process of introgression, which marks the gene flow of the interspecific hybrid to the parent species through backcrossing. In this way the parent species may become genetically 'swamped' by the introgressed genes from the other species (Secondi et al. 2006). Furthermore, an imbalance in the fertility between both sexes of hybridized offspring is a common occurrence in which the heterogametic sex is not as fertile or sometimes sterile when compared to the homogametic sex. This phenomenon is known as Haldane's rule. In birds, females are the heterogametic sex, therefore Haldane's rule is a topic of particular interest in avian studies because females are also the dispersing sex (Pons et al. 2014). This fact coupled with the maternal

inheritance of mitochondrial DNA may lead to differential introgression of autosomal, sex-linked, or mitochondrial loci in cases of hybridization where Haldane's rule applies (Carling and Brumfield 2008). It is also worth mentioning that Haldane's rule does not apply to every case of hybridization and operates on more of a continuum by it and exceptions to the rule are also common (McCarthy 2006).

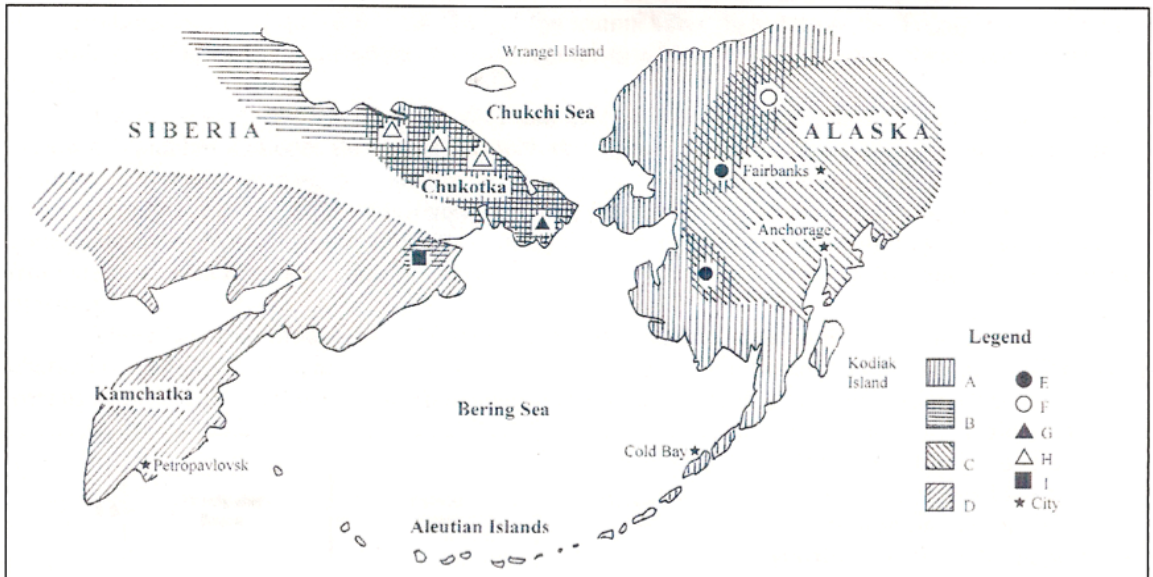

Figure 1: Breeding areas [with overlaps] of Northern Swans in Siberia and Alaska
A = Tundra; B = Bewick's; C = Trumpeter; D = Whooper; E = known overlap of Trumpeter and Tundra; F = possible overlap of Trumpeter and Tundra; G = known overlap of Bewick's and Tundra; H = probable overlap of Bewick's and Tundra; I = known overlap of Bewick's and Whooper. Compiled from (Degernes et al. 2002; Syroechkovski 2002)

Figure 2. Breeding Range overlap of Swans in the genus *Cygnus*. (Sladen and Gillevet, 2007)

Conservation Efforts in Overlapping Ranges

The Trumpeter swan (*Cygnus buccinator*) and the Tundra swan (*Cygnus columbianus*) share many phenotypic traits to the extent that they are nearly non-distinguishable in the wild (Banko 1960). While *Cygnus buccinator* was nearly extirpated by over-harvesting in the late 19th century, the populations of Trumpeters rebounded after the establishment of the Migratory Bird Treaty Act of 1918 followed by protected refuges in designated areas. Climate change and anthropogenic disturbances such as habitat destruction have led to a pattern of northward movement of the Trumpeter swan into the breeding zones of Tundra swans (Beck et al. 2011, Schmidt et al. 2011, Schmidt et al. 2009,). This is a concern for the conservation of Trumpeter Swans for two reasons. Firstly, Trumpeter swans could be mistaken for Tundra Swans by hunters and may become accidental victims (Engelhardt 2000). The second reason for concern is derived from the well-documented ability of both species to form fertile hybrids (Travsky and Beauvais 2004). As mentioned in the previous section, if hybrids are being formed then there is risk of 'swamping' out the genetic material of each species through introgression.

Over recent years, the Swan Research Program at Airlie has focused its efforts on developing methods of detecting wild Trumpeter-Tundra hybrids. This work has involved the collection of Trumpeter and Tundra swan samples from the

Koyukuk National Wildlife refuge, Alaska and the subsequent comparison of such samples with hybrids reared in captivity. The Micbrobiome Analysis Center (MBAC) at George Mason University has been heavily involved in the molecular and genetic analysis portion of this work. To distinguish Trumpeters from Tundra Swans, as well as Trumpeter-Tundra hybrids from their parent species, Lauren Wilson at MBAC developed and tested new microsatellite molecular markers derived from Next-gen sequencing runs obtained using the genomes of both species as a template. Additionally, Elizabeth Dingess at MBAC analyzed the sex-linked chromo-helicase-DNA binding protein encoding gene *CHD* and its associated introns, A and E, in Trumpeters, Tundra Swans, and their hybrids. This study takes the research of hybridization between swan species further by placing the impetus on examining the genetic relatedness of the Trumpeter Swans and Tundra Swans rather than on the development of the novel genetic markers for introgression.

Study Aims

The aim of this study was to examine the extent of genetic relatedness between *Cygnus buccinator* and *Cygnus columbianus columbianus* by a genomic comparison between the two species.

To this end, our plans included an assembly of Trumpeter and Tundra raw genomic reads with the use of various bioinformatics software. Typically,

genomic assembly involves forming scaffolds and generating consensus sequences from the Contigs (Contiguous sequences of DNA). In this study, custom Perl scripts were developed in order to print out summary statistics for each Contig.

Our plans also included further analysis of the genetic relatedness between swan species using sequences of nuclear gene encoding chromo-helicase-DNA binding protein (*CHD)* and associated introns that involved alignments leading to phylogenetic tree construction. Additionally, polymorphisms in the mitochondrial control region (D-loop) of Trumpeters, Tundra Swans, and hybrids were used for phylogenetic trees analysis. The trees obtained using the nuclear gene *CHD* and D-loop were compared to each other.

Figure 3a. Tundra Swan (*C. colubianus columbianus*) T185 found on the Colville River Delta in North Slope, Alaska.

Figure 3b. Trumpeter (*C. buccinator)* 13UK found in Koyuku Wildlife refuge, Alaska.

MATERIALS AND METHODS

Sample collection and sequencing

Genomic DNA samples

Two DNA samples (one representative *C. buccinator* and one *C. columbianus columbianus*) were used for Next-Generation sequencing. Both species of swans were identified based on morphology and vocalizations by members of the Arlie Swan Research Center. The individual swans were subsequently banded, and blood samples were drawn in the field. Tundra T185 (Figure 3a), an adult male, was found on the Colville River Delta in the North Slope of Alaska (Figure 4) on June 27, 2006. Trumpeter 13UK (Figure 3b), an adult female, was found on the Koyukuk National Wildlife Refuge and banded in August 2006 (Figure 4). Sequencing of Trumpeter 13UK and Tundra T185 was performed on a Roche 454 GS-Junior by staff at the Microbiome Analysis Center, George Mason University.

Sanger Sequencing of Nuclear Gene CHD

The sex-linked chromo-helicase-DNA binding (*CHD*) gene and its associated introns, A and E, of Trumpeter and Tundra Swan samples from different sampling locales (North Slope and Kaiyuh, Alaska) were sequenced at

the Microbiome Analysis Center at George Mason University. In addition to the specimens of known species origin, the same gene was sequenced in twenty-four (24) specimens of wild swans collected in Washington State with unknown species and gender identification by Elizabeth Dingess. These specimens were obtained from the Airlie Swan Research Program. Both the Z and W chromosomes of Intron A were amplified with allele-specific primer sets designed by Dingess (2008, unpublished data).

Sanger Sequencing of Mitochondrial DNA

Sequencing of the mitochondrial control region (D-loop) of *Cygnus buccinator, Cygnus columbianus columbianus* and their hybrids was performed by the Microbiome Analysis Center. The sequencing included Trumpeters and Tundra swans representative of several different localities in Alaska, which were Kaiyuh Flats, Koyukuk, North Slope, and Selawik. F1 and F2 generation hybrids (Trumplings) in addition to backcrosses (hybrid x Trumpeter or Tundra) from Airlie Swan Research Center were included in the sample set. A total of forty-two (42) Trumpeter swans, sixteen (16) Tundra swans, one (1) Trumpetling, and four (4) Trumplings had their mitochondrial D-loop sequenced.

Figure 3. Sampling locations for Trumpeter 13UK and Tundra T185 used in genomic analysis.

Analysis of Genomic Sequences

In comparison to traditional capillary sequencing methods, NextGen Sequencing methods produce an enormous volume of data at a reduced cost (Lerner and Fleischer 2010; Metzker 2010). In NGS, the clonal amplification step involves replication of fragmented template DNA that is carried out and

subsequently sequenced in parallel on a substrate that is unique to each platform (Lerner and Fleischer 2010). NGS is often performed where there is little amount of material available or where the number of individuals sampled from a population is limited.

In typical Next-Generation sequencing attempt, short reads of around 400 basepairs long are produced and assembled into contigs, or contiguous overlapping segments, by the software accompanying the respective platform. These paired-end contigs are further assembled with gaps of known length into longer regions known as scaffolds. Scaffolds are then mapped to a reference genome or assembled by de novo alignment, without a reference.

In addition to relatively large amounts of sequencing errors, a major challenge for NGS sequencing in general is posed by repeats in the genome of the organism of interest. These repetitive elements make de novo alignment or mapping the sequenced scaffolds to reference genome difficult as they align to multiple repetitive locations on the reference. Because these repeats are not exact copies of each other due to variations in their sequence, the reference alignment will have different match scores at the repeat locations (Treangen and Salzberg 2012). Additionally, repeats introduce the gaps in the assembly at locations when the repeat length is longer than the read length (Treangen and Salzberg 2012).

Despite these drawbacks of NGS, the avian genomes in comparison to other vertebrate genomes are a good target for *de novo* assembly because they are smaller in size and have substantially lesser proportion of repeats. For instance,

Hughes and Piontkivska (2005) found that only 10.3% of the chicken genome is occupied by repeats in contrast to the 45% of the human genome occupied by repeats. For this reason coupled with the absence of a reference genome for the two swan species of study (Tundra and Trumpeter), we performed the sequencing followed by *de novo* assembly.

Trumpeter and Tundra raw reads were assembled *de novo* in the sequence editing and analysis package, Geneious 7.0.6 (Biomatters Ltd). The FASTA files for Tundra and Trumpeter data were assembled to each other at 80% Overlap Identity. As we were specifically interested in finding the differences in two swan species, the species-tagged reads from both species were used in the same assembly. As defined in Geneious 7.0.6 software manual, the overlap identity is the minimum percent sequence identity of the overlap region between a sequence and any sequence in the contig required for the sequence to be included.

FASTA files for the resulting Contigs were compared and processed with a custom Perl script referred to as the Contig Comparison script (see appendix). The script was designed to enumerate differences in sequence identity between the two species; Trumpeter and Tundra, within each Contig. The Contig Comparison script first parsed the sequences based on species identity and then compared those sequences within each contig with respect to the species they belonged to. The amount of matches, mismatches, SNPs and GC content were printed out in summary statistics for each file. The script was specifically designed to look at the overlapping regions of sequence between the two species

and disregard overhangs and gaps. Because of that, the input sequence trimming was not required. A match was called if for a particular base all sequences for both species were identical and there were no gaps in either sequence. In the same way, a mismatch was determined if the bases at a particular position for respective species were not the same. Variants were ascertained by determining the instances where the IUPAC codes of the two species indicated the presence of polymorphism. For instance, if at a certain position a nucleotide from a Trumpeter sequence was C or G (IUPAC code = S) and at the same position a nucleotide from a Tundra sequence was a G, the script identified a variant present at this location. In this regard, a SNP is the difference within either Tundra or Trumpeter–specific Contig within either Tundra Swans or Trumpeters and a mismatch is a difference between the two species. The percent of G and C bases in each contig were also calculated. Coverage was approximated with the Lander/Waterman equation (Lander and Waterman, 1988). Coverage= LN/G, where L= read length, N = number of reads, and G equals the target genome length. Since we don't know the size of the genomes of *C. buccinator* or *C. columbianus columbianus,* the coverage was calculated for G equal to the smallest avian genome 531.96 Mb (*Coturnix japonica* NCBI ID: 113) and the largest avian genome 1548.48 Mb (*Aquila chrysaetos* NCBI ID: 32031) available from NCBI. Coverage was calculated for the reads belonging to the Trumpeter and Tundra Swan in addition to the total overlap coverage between the two species.

Analysis of the sequences of the nuclear gene CHD

FASTA reads from the CHD nuclear gene on the W chromosome of seventeen (17) swan specimens with unknown species identification (labeled SRP) and five (5) specimens of known species origin (2 Tundra and 3 Trumpeter Swans) were imported into Geneious 7.0.6. The Sequences were aligned in Geneious 7.0.6 using a MUSCLE (Multiple Sequence Comparison by Log-Expectation) alignment under default parameters. MUSCLE is a multiple alignment that uses both *k*-mer distances and Kimura distances to align sequences. The MUSCLE alignment method has the advantage of speed in performance when compared to similar alignment methods (Edgar 2004). Three sequences; SRP 20_AW_32, SRP 13_CHD_57, and SRP 12_CHD_56 were removed from the alignment because they were duplicate sequences that were deemed poor quality as a result of sequencing error. The resulting alignment was trimmed to a length of 399 basepairs. The alignment was used to build a Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Intron A of the W allele of the CHD nuclear gene (Figure 5) in Geneious 7.0.6.

FASTA reads from the CHD on the Z chromosome were analyzed with the same methods used in the analysis of CHD-W as mentioned above. Sequences from sixteen (16) swan specimens with unknown species identification (labeled SRP) and eight (8) morphologically identified specimens (4 Tundra and 4

Trumpeter Swans) were imported into Geneious 7.06. Nine SRP sequences were removed from the analysis based on poor quality as shown by their fluorescent chromatographic profiles. The resulting nine SRP sequences, 4 Trumpeter and 4 Tundra swan sequences were aligned and the alignment was used to build the Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Intron A of the Z allele of the CHD nuclear gene (Figure 6).

Analysis of the sequences of the mitochondrial DNA (D-loop)

FASTA reads for the D-loop of the 63 samples were imported into Geneious 7.0.6 for editing, alignment, and SNP analysis. Sequences were aligned using a MUSCLE alignment in Geneious 7.0.6 under default parameters. Sequences from the mtDNA D-loop of Trumpeter R31A, Tundra T213, and Trumpling Z363 were removed from the alignment because they were deemed poor quality based on the fluorescent chromatographic profiles.

A neighbor-joining phylogenetic tree (Figure 6) of the alignment was constructed using the Geneious Tree Builder in Geneious 7.0.6. Variants in the sequences were searched for using the default parameters in Geneious 7.0.6 with a Minimum Variant Frequency of 0.25. Both inter-species (between species) SNPs and intra-species (within species) SNPs variants were selected and enumerated with the same default parameters. Loci between species SNPs were exported in FASTA format and imported into GenAlEx 6.501 (Peakall and Smouse 2006,

2012) for conversion of nucleotide SNPs into integers 1-5 (1=A, 2=T, 3=G, 4=C, 5=missing data). GenAlEx 6.501 is population genetics analysis software that runs in Microsoft Excel. The integers were used as input data for the statistical program STRUCTURE *2.3.4* to cluster the 60 individual swans into different probable populations based on SNPs. STRUCTURE implements a Bayesian clustering procedure to probabilistically assign individuals to populations. Markov Chain Monte Carlo (MCMC) process was employed to estimate the number of populations (K) and assign individuals in the sample to K populations to achieve Hardy-Weinberg equilibrium (Falush *et al.* 2003; Pritchard *et al.* 2000). Based on the likelihood data-plot of the simulation runs, a burn-in of 20,000 with a run-length of 60,000 steps was chosen.

Software Used

1. **Geneious 7.0.6** Biomatters Ltd. http://www.geneious.com/
2. **GeneAlEx 6.5** Peakall, R. and Smouse P.E. (2006) GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. Molecular Ecology Notes. 6, 288-295. http://biology.anu.edu.au/GenAlEx/Download.html
3. **STRUCTURE *2.3.4*** Inference of population structure using multilocus genotype data. J.K. Pritchard, M. Stephens and P. J. Donnelly, 2000. Genetics 155: 945-959. http://pritchardlab.stanford.edu/structure_software/release_versions/v2.3.4/html/structure.html

Table 1. Summary Statistics generated by Compare Contigs Script.

Differences	**5180**
Matches	**221033**
Mismatches	**6815**
SNPs	**1635**
Total Positions	**227848**
Average Percent Match	**97.01%**
Average Percent Mismatch	**2.99%**
Average percent SNPs	**0.72%**
Average Percent Difference	**2.27%**

RESULTS

Analysis of the Genomic Reads obtained using Trumpeter and Tundra Swan DNA samples as a template

Shotgun sequencing resulted in a total of 91,840 reads for Trumpeter Swan, and a total of 103,942 reads for Tundra swan. Assembly of the sequence reads produced a total of 572 contigs. The consensus contigs were aligned and subsequently run through the Contig Comparison script that compares the overlapping sequences from the Trumpeter and Tundra in the Contig. The summary statistic output of the Contig Comparison script is represented in Table 1. The average percent match between the Contigs built using reads from both species, Trumpeter and Tundra, was 97.73% with a standard deviation of 2.9%. The average GC content in the contigs was at 46.31%. The coverage for Trumpeter and Tundra, assuming the genome sizes were comparable to the size of the smallest avian genome (531.96 Mb) belonging to *Coturnix japonica* NCBI ID: 113, was 9.29% and 10.51% respectively. The overlapping coverage of a genome this size would be 0.04%. Assuming that the total length of the swan genomes is close to the largest avian genome deposited in NCBI (1548.48 Mb), the coverage would be at is 3.61% for both Trumpeter and Tundra species and 0.01% for overlapping sequences.

Analysis of the sequences of the nuclear gene CHD

The alignment of 32 CHD-W sequences (27 SRP sequences, 3 Trumpeter sequences, and 2 Tundra sequences) had a pairwise percent identity of 92.7%. The average un-gapped length was 375 basepairs with a minimum of 270 basepairs and a maximum of 393 basepairs. The frequency of adenine was 36.1%, the frequency of cytosine was19.8%, while the frequencies of guanine and thymine were 14.2% and 29.9%, respectively, with overall GC content at 32%. In the sequences of two Tundra Swans (Tundra 214 and Tundra 727) and two of the unknown Swans (SRP 12 and SRP13), there were large deletions with a size of 126 basepairs. A total of 28 sequences (3 Trumpeters and 25 SRP sequences) without the 126 basepairs deletion demonstrated a 99.6% pairwise identity. In the Neighbor-Joining phylogeny that was built using alignment of the 32 CHD-W sequence (Figure 5), all Trumpeters, Tundras and 23 SRP samples form a distinct clade, while samples SRP 16 and SRP 22 form a monophyletic group that places outside of the majority of the samples, while samples SRP 25 and SRP 11 are classified as paraphyletic.

The alignment of 15 CHD-Z sequences (7 SRP sequences, 4 Trumpeter sequences, and 4 Tundra sequences) had a pairwise percent identity of 99.5%. All sequences had a length of 467 basepairs. The frequency of adenine was 31.5%, the frequency of cytosine was 20.6%, while the frequencies of guanine and

thymine were 20.8% and 26.9%, respectively, with overall GC content at 41.5%. In the Neighbor-Joining phylogeny that was built using alignment of 15 CHD-Z sequences (Figure 6), Tundra T228 and Tundra 727 form a monophyletic group and Trumpeters R05 and R06 form a separate clade with SRP 6. In contrast, the other Trumpeters; R12 and R42, and two Tundra Swans; 972 and T221 are grouped together in the same clade.

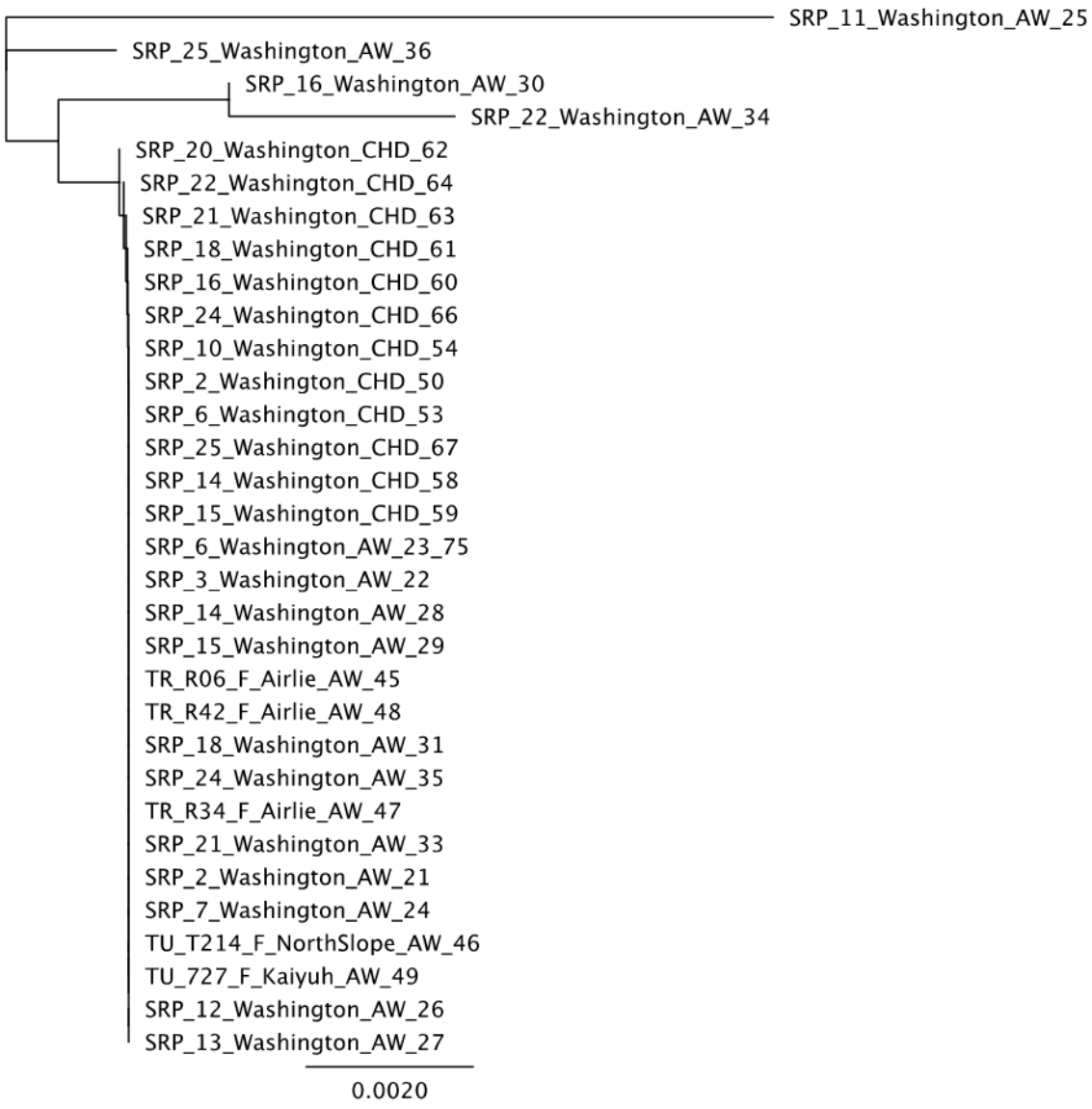

Figure 5. Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Intron A of the W allele of the CHD nuclear gene.

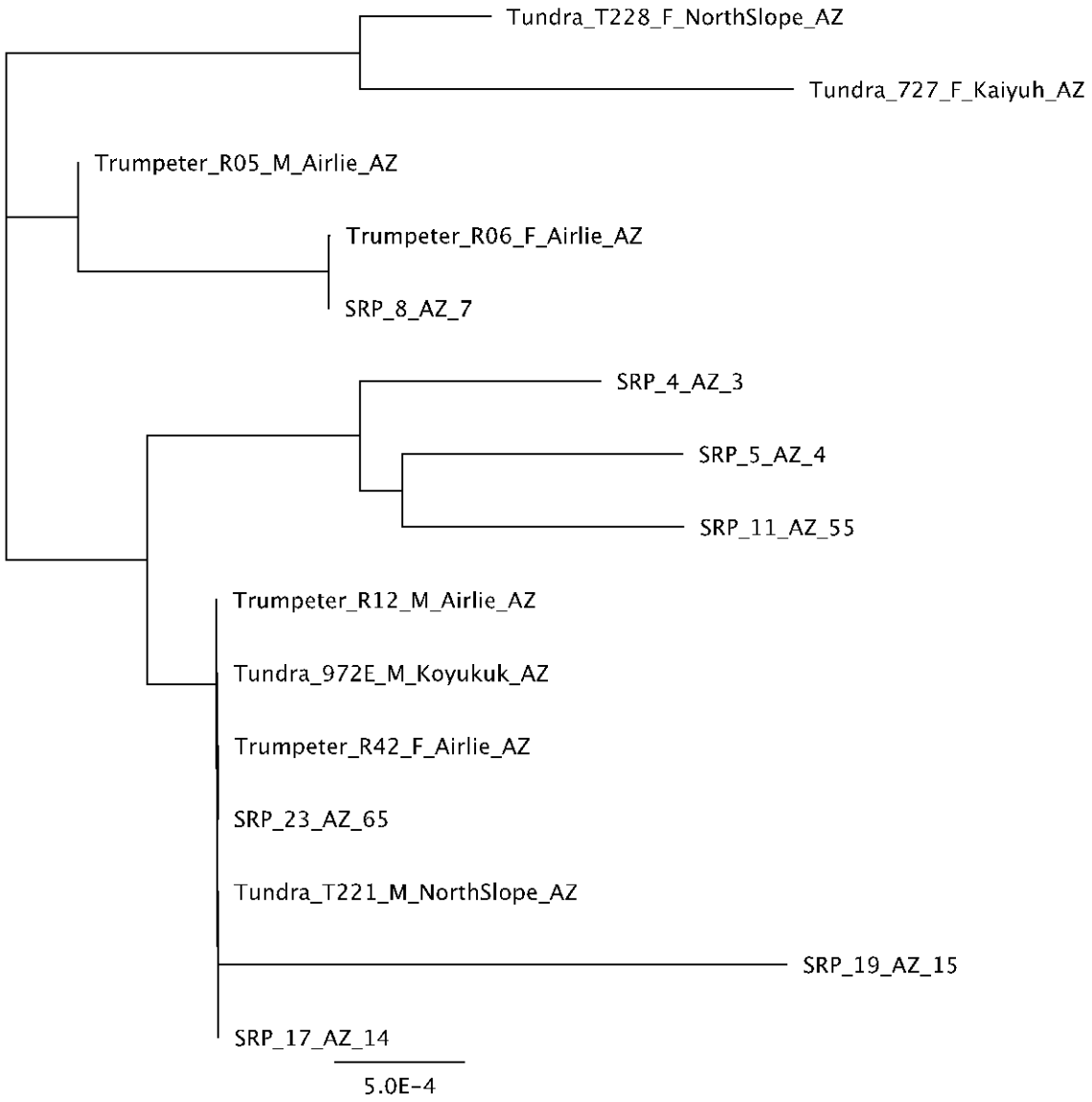

Figure 6. Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Intron A of the Z allele of the CHD nuclear gene.

Analysis of the Mitochondrial (D-loop) Sequences

A majority of the mtDNA D-loop sequences of 60 Trumpeter, Tundra and hybrid were 1098 basepairs long, while a few sequences were a bit shorter, 978 basepairs, 1033 basepairs, and 1061 basepairs in size. Across all samples, there were 46 SNPs, 42 of which were transitions (purine to purine or pyrimidine to pyrimidine) and 4 of which were transversions (purine to pyrimidine or *vice versa*). A majority of interspecies variants were found in between nucleotide positions 1 to 250 and 690 to 1098, while the locus between the positions 317 and 784 was least polymorphic. Species-specific analysis of D-loop variants resulted in two (2) SNPs detected in Trumpeter swans and ten (10) SNPs detected in Tundra swans.

The Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Mitochondrial DNA (D-loop) (Figure 7) shows that Trumpeter swans form a clade that is distinct from Tundra Swans, while four Trumpetling D-loop samples (Trumpetlings Z380A and B, Z557A and B) and three Trumpling D-loop samples (Z595, Z531, and trumpling Z377) are branched with Trumpeters. A duplicate sequence of Trumpling Z377 is conflicting with trumpling Z377 in its placement outside of both clades. Trumpling Z321, an F2 hybrid, groups with the Tundra Swans.

The results from STRUCTURE *2.3.4* (Figure 8) suggest that the most likely partitioning of the data exists for K=2, or two clusters. There are some conflicting results regarding the clustering of hybrid swans, specifically with

respect to specimens Z377 and Z531. A specimen Z531 was collected from a Trumpling (Individual #53) that clusters more strongly with the Tundras according to its mitochondrial DNA sequence. Neighbor-Joining phylogeny according to the sequences of the Mitochondrial DNA (D-loop) (Figure 7) places Trumpling Z531 closer to the Trumpeters than the Tundra Swans.

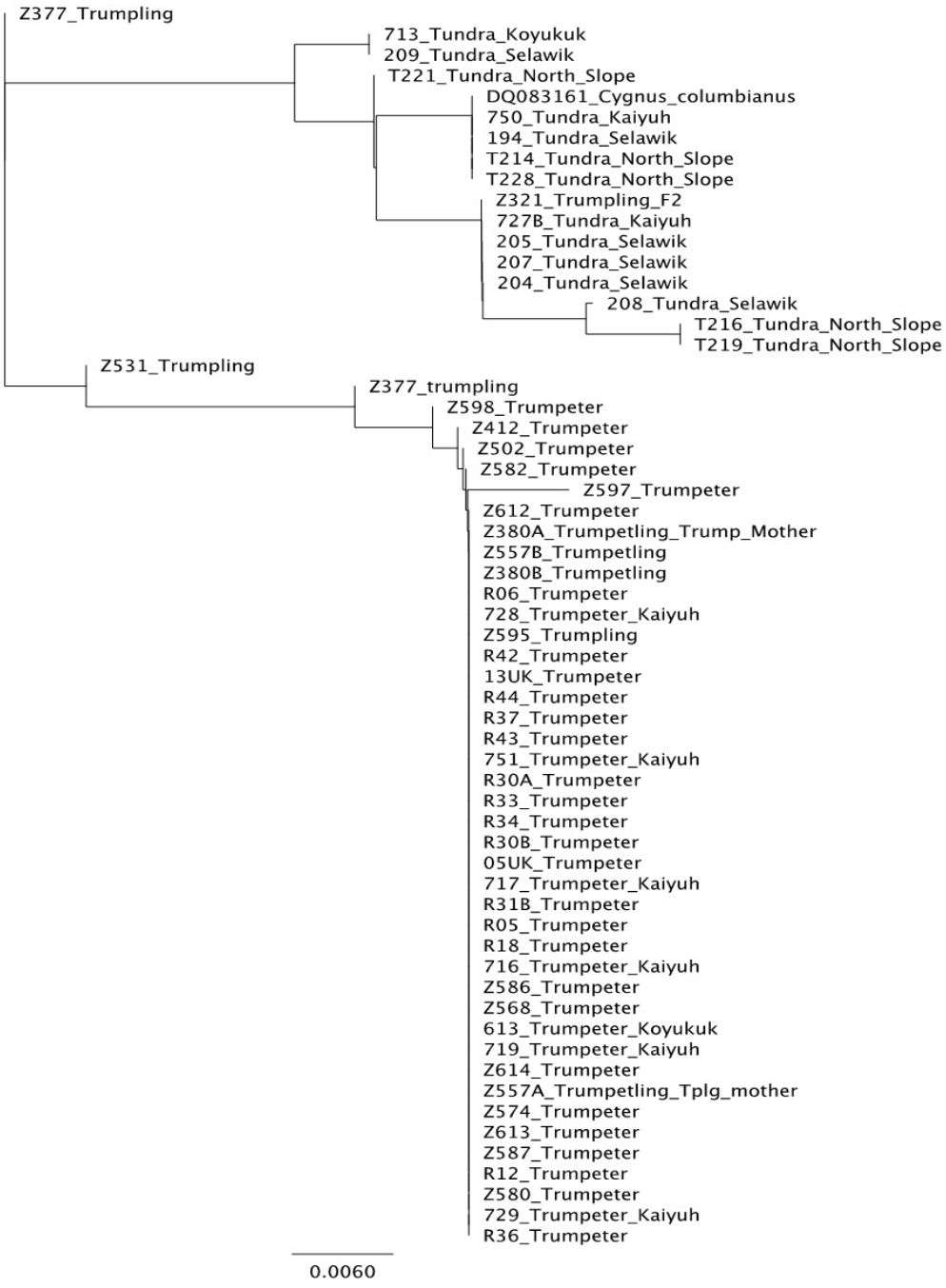

Figure 7. Neighbor-Joining phylogeny of Trumpeter, Tundra Swans and hybrids according to the sequences of the Mitochondrial DNA (D-loop).

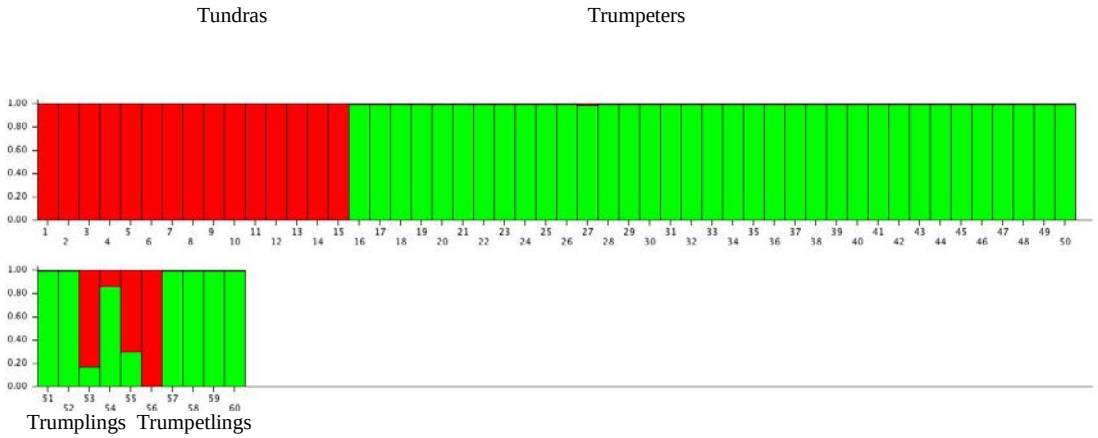

Figure 8. STRUCTURE Barplot analysis of mtDNA (D-loop) data for Tundras (1-15), Trumpeters (16-52), and Hybrids (53-60). The Barplot represents results of 2 genetic clusters (K=2). Each column represents an individual swan and each color represents a cluster. The extent of color represents the estimated membership coefficient of the individual in that cluster.

DISCUSSION

Our comparison of *Cygnus buccinator* and *Cygnus columbianus columbianus* using two different types of DNA (nuclear and mitochondrial) reveals some major differences in the ability of each type of data to distinguish the species origin of the collected specimens. The Analysis of the Genomic Reads obtained using Trumpeter and Tundra Swan DNA samples as a template demonstrates how remarkably close the species are despite some of the key morphological differences such as vocalization, bill pattern, and size differences. Importantly, the mtDNA profiling was able to clearly identify the two species.

The assembly of sequence reads from the two samples revealed a high degree of similarity between two swans profiled, Trumpeter 13UK and Tundra T185. A difference between two species was, on average, 2.27% in sequence of the contiguously aligned fragments. This high degree of similarity was surprising even considering the morphological resemblances of Trumpeter and Tundra Swans. Given that most of the completely sequenced genomes of avian species are between 531.96 Mb (*Coturnix japonica* NCBI ID: 113) and 1548.48 Mb (*Aquila chrysaetos* NCBI ID: 32031) in size (NCBI), we estimate that we aligned between 0.01 and 0.04% of the Trumpeter and Tundra swan genomes which is orders of magnitude smaller than the actual genome of each species. However,

since Next-Gen sequencing contigs evenly cover genomes, we can assume that the sampling is representative of the total genome.

The chromo-helicase-DNA binding (*CHD*) is a relatively recently described nuclear gene located on the W chromosome and is suitable for determining the sex and species of birds. In all birds, females are the heterogametic sex carrying both Z and W chromosomes while males carry two Z chromosomes. We analyzed intron A of this gene in morphologically identified Trumpeters and Tundra swans and in a number of specimens of uncertain species origin. The Neighbor-Joining phylogenetic analyses of *CHD*-W and *CHD*-Z of intron A (Figure 5 and 6, respectively) corroborate the high degree of similarity found in the genomic comparison between the two species. This is suggested by the grouping of two known Tundra Swans; Tundra 214 and Tundra 727 from two distinct geographic locales with three known Trumpeters; Trumpeter R06, Trumpeter R42, and Trumpeter R34 from Airlie into one clade on the Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Intron A of the W allele of the CHD nuclear gene (Figure 5). Interestingly, both species were grouped into the same clade despite the presence of a large 126 basepairs deletion between nucleotide positons of 40 and 166 within the CHD-W gene found in the Tundra swans and two samples of unknown origin (SRP 12 and SRP13).

The Neighbor-Joining phylogeny of SRP (unknown species), and known Trumpeter and Tundra Swans according to the sequences of the Intron A of the Z

allele of the CHD nuclear gene (Figure 6) further supports the genetic similarity of the species in its inability to discriminate between four morphologically identified Trumpeters and Tundra Swans. This lack of differentiation between Trumpeter and Tundra Swans is evident from the grouping of two known Trumpeters; R12 and R42 with 2 known Tundra Swans; T221 and 972.

Contrasting the high degree of similarity demonstrated in the analysis of the nuclear gene (*CHD*) and the genomic reads of Trumpeter and Tundra swans, the Mitochondrial (D-loop) sequences clearly distinguish Trumpeters from Tundra swans. As could be noted in the alignment of the mtDNA (D-loop) sequences (data not shown), two major mitogroups are apparent; one for Trumpeters and one for Tundra Swans. Hence, it was not surprising that the Neighbor-Joining phylogeny of Trumpeter, Tundra Swans, and hybrids according to the sequences of the Mitochondrial DNA (D-loop) (Figure 7) clearly separates Trumpeters from Tundra Swans into two distinct clades. As evident from the tree, the grouping of hybrids into the mtDNA based clades follows maternal inheritance. For instance, Trumpetling Z380 groups with Trumpeters despite its father being a Trumpeter/Tundra hybrid, while the Swan's mother was a Trumpeter. There is some discrepancy with the placement of hybrid Z377 on the tree, but as previously stated, this conflict can be attributed to sequencing error. Given the historical context of the bottleneck in Trumpeter populations, results of the analysis of the mtDNA D-loop look interesting. Although a presence of bottleneck was not explicitly tested for, the within-species SNP analysis of both

species showed there to be five times the amount of variance in the D-loop of Tundra Swans as compared to the D-loop of Trumpeters (10:2, respectively). Greater genetic diversity in the Tundra swans is also evident from the Neighbor-Joining D-loop–based tree that is better structured and has longer divergence times in the branching of the Tundra Swan clade.

The ability of the mtDNA to differentiate the species specificity was also supported by the D-loop SNP analysis carried out in STRUCTURE 2.3.4. (Figure 8). Results from STRUCTURE 2.3.4 corroborate with the phylogenetic analysis of the D-loop. The best representation of the data was met with K=2 or two populations.

The difference in the trees generated by the mitochondrial DNA and nuclear DNA could reflect the type of introgression between Trumpeters and Tundra Swans. Given how common hybridization and introgression are in avian interactions, this is not unexpected for this type of dynamic to have occurred in Trumpeters and Tundra Swans, especially in areas of overlapping breeding ranges. Furthermore, varying rates of introgression dependent on the type of DNA (mitochondrial or nuclear) may be detected (Pons et al. 2013). This discordance between mtDNA introgression and nuclear DNA introgression can be so great that in certain cases one of the named type may be non-detectable (Pons et al. 2013). Specifically, higher rates of introgression of nuclear genes in comparison to mitochondrial are well documented in birds (Petit and Excoffier 2009; Pons et al. 2013). It is interesting to consider the results obtained in this study in

comparison to those of Oyler-McCance et al. (2007) in which two distinct populations of Trumpeters (Pacific Population and Rocky Mountain Population) were identified. One could speculate that the genetic differentiation found by Oyler-McCance et al. (2007) stems from differential hybridization, in which one population inside the contact zone of the two species experiences recurring hybridization events, while the other population outside of the contact zone remains conspecific in breeding. However, this theory would have to be tested with wild Trumpeter Swan samples from distinct locations in the ecotone (the area where the two ranges overlap) and outside of the ecotone.

SUMMARY

In this study, a genetic comparison of the Trumpeter Swan (*Cygnus buccinator*) and the Tundra Swan (*Cygnus columbianus columbianus*) was performed. The high amount of genetic similarity between the two species revealed in analysis of their genomic sequences in addition to sequences of nuclear gene encoding chromo-helicase-DNA binding protein (*CHD*) was in stark contrast to the ability of the mitochondrial DNA (D-loop) to clearly distinguish the species. This discordance between the species-differentiating properties of mtDNA and nuclear DNA suggests that the introgression between the Trumpeter Swans and Tundra Swans is a result of extensive hybridization. Further analysis using the same methods with a larger sampling from areas within the range overlap of the two species and areas outside of the ecotone should be carried out to validate this finding.

APPENDIX

Perl Script for Contig analysis

```
#!/usr/bin/perl
# Copyright 2013 P.M. Gillevet

use strict;
#use warnings;

#  12/21/2013   Compare Swan Contigs from Genious and Tabulate results
#  01/02.2014     Added GC Content          (Chris Yesmont)

#####################################################
print "\nFORMAT:  Compare_Swan_Contigs.pl Alignment_File.txt\n\n";

my $fasta_file = @ARGV[0] || die;
system ("perl -pi -e 's/\f//g' $fasta_file");
system ("perl -pi -e 's/\r/\n/g' $fasta_file");

my $outputfile;
$_ = $fasta_file;
if (s/\..*/\_Results\.txt/) {$outputfile = $_;}
print "Output file is $outputfile \n";
open (RESULT,">$outputfile");

my $outputfile2;
$_ = $fasta_file;
if (s/\..*/\_Summary\.txt/) {$outputfile2 = $_;}
print "Summary file is $outputfile2 \n";
open (SUMMARY,">$outputfile2");

#################################################
print "Parse fasta_file\n";

my $handle2 = andopen_file($fasta_file);
my $file2 = $handle2;

my $line;
my $sequence;
my $name;
my $file;
```

```
my @Tundra;
my $Tundra;

my @Trumpeter;
my $Trumpeter;

my @Sequence;
my $Sequence;

my $i;

while ($line = <$file2>)
       {
            chomp $line;

        if  ($line =~ /\>/)
           {
                $line =~ s/\>//;

                ($name) = split(/_/,$line);
                     $sequence = "";
                     }

            else
               {
           @Sequence = split(",$line);

                        if ($name =~ /tundra/i)
                          {
                          foreach ($i=0; $i<@Sequence;$i++)
                   {

                                  if ($Tundra[$i] eq "" or $Tundra[$i] eq "-" ) {$Tundra[$i] =
@Sequence[$i];}

                        }
                          }


                        if ($name =~ /trumpeter/i)
                          {
                          foreach ($i=0; $i<@Sequence;$i++)
                   {

                                  if ($Trumpeter[$i] eq "" or $Trumpeter[$i] eq "-" )
{$Trumpeter[$i] = @Sequence[$i];}

                        }
                          }
                }
```

```
        }

########################## compare Tundra and Trumpeter

my $count;
my $match;
my $mismatch;
my $SNP;
my $Diff;
my $GC;

# Define IUPAC Codes
my $IUPAC;

my %IUPAC = (
A => '[A]',
C => '[C]',
G => '[G]',
T => '[T]',
R => '[AG]',
Y => '[CT]',
M => '[AC]',
K => '[GT]',
W => '[AT]',
S => '[GC]',
B => '[CGT]',
D => '[AGT]',
H => '[ACT]',
V => '[ACG]',
N => '[ACGT]',
);

foreach ($i=0; $i<@Tundra;$i++)
  {

   $count++;
   if ($Tundra[$i] eq $Trumpeter[$i] and $Tundra[$i] ne "-" and $Trumpeter[$i] ne "-" )
{$match++;}

   if ($Tundra[$i] ne $Trumpeter[$i] and $Tundra[$i] ne "-" and $Trumpeter[$i] ne "-" and
$Tundra[$i] ne "" and $Trumpeter[$i] ne "")
      {$mismatch++;
       print RESULT "Mismatch\t$fasta_file\t$i\t$Tundra[$i]\t$Trumpeter[$i]\t$mismatch\n";
      }

   if ($IUPAC{$Trumpeter[$i]} =~ $IUPAC{$Tundra[$i]} and $Trumpeter[$i] ne
$Tundra[$i])
       {
           $SNP++;
            print RESULT "SNP\t$fasta_file\t$i\t$Tundra[$i]\t$Trumpeter[$i]\t$SNP\n";
```

```
            }

    if ($IUPAC{$Trumpeter[$i]} =~ "C" or $IUPAC{$Trumpeter[$i]} =~ "G" and
$Tundra[$i] ne "-") ###### calculate GC content (Chris Yesmont 01/02/2014)
        {
            $GC++;
            }


  }

########################  print ut summary stats
print RESULT "\n\n";

my $total = $match + $mismatch;
my $Match_Percentage = $match/$total;
my $Mismatch_Percentage = $mismatch/$total;
my $SNP_Percentage = $SNP/$total;
my $difference = ($mismatch-$SNP);
my $total_difference = ($mismatch-$SNP)/$total;
my $total_similarity = 1-$total_difference;
my $GC_Content = $GC/$total;


print SUMMARY
"File\tTotal_Length\tPositions_Compared\tMatches\tMisMatches\tSNPs\tDifference\tMatch_
Percentage\tMisMatch_Percentage\tSNP_Perentage\tTotal_Difference\tTotal_Similarity\tGC
_Content\n";
print SUMMARY "$fasta_file\t";
print SUMMARY "$count\t";
print SUMMARY "$total\t";
print SUMMARY "$match\t";
print SUMMARY "$mismatch\t";
print SUMMARY "$SNP\t";
print SUMMARY "$difference\t";


print SUMMARY "$Match_Percentage\t";
print SUMMARY "$Mismatch_Percentage\t";
print SUMMARY "$SNP_Percentage\t";
print SUMMARY "$total_difference\t";
print SUMMARY "$total_similarity\t";
print SUMMARY "$GC_Content\n\n";                          #print out GC content

############
close (RESULT);
close (SUMMARY);


exit;
```

```
##############################################
##############################################
#SUBROUTINES
####################################
####################################
#print array

sub print_array
{

my(@array);

 @array = @_;
 foreach (@array)

       {
         $line = $_;
              print $line;
              print "\n";

              }
return;
}

#####################################
# open_file
#
#   - given filename, set filehandle

sub open_file {

   my($filename) = @_;
   my $fh;

   unless(open($fh, $filename)) {
      print "Cannot open file $filename\n";
      exit;
   }
   return $fh;
}


########################################
#  Print Hash

sub print_hash
{
my (%hash)= @_;
my ($key);
my ($value);
```

```
while (($key, $value) = each %hash)
    {
     print "$key = $value";
     print "\n";
     }
return;
}



########################################
#  Print sorted Hash by value
# flip key and values then sort

sub print_sort_value
{
my (%hash)= @_;
my ($key);
my ($value);

my (%flipped_hash);
my ($flipped_hash);
my (@sortorder);

while (($key, $value) = each %hash)
    {
     $flipped_hash{$value} = $key;
     }

@sortorder = sort keys %flipped_hash;

foreach $value (@sortorder)
    {
     print $flipped_hash{$value}, " = ",$value ;
     print "\n";
     }
return;
}

########################################
#  Print sorted Hash by key

sub print_sort_key
{
my (%hash)= @_;
my ($key) = '';
my ($value) = '';
my (@sortorder);
```

```
@sortorder = sort keys %hash;

foreach $key (@sortorder)
    {
     print $key, " = ", $hash{$key};
     print "\n";
     }
return;
}


################################################################################
#####
##  Tabulating the count of an array

sub tabulate

{
my (@array) = @_;
my %tabulate;
my $tabulate;
my $array;

for (@array)
   {
        $tabulate{$_}++;
   }

return %tabulate;
}
```

REFERENCES

Baldassarre, G.A., and Bolen, E.G. (1994). Waterfowl Ecology and Management. New York, New York: J. Wiley.

Banko, Winston E. (1960). The Trumpeter Swan Its History, Habits, and Population in the United States. North American Fauna, 63, 1-214.

Banko, W.E., Schroger, A. W. (1976). Handbook of North American Birds 2. New Haven, Connecticut: Yale University Press.

Barrett, V. A., and Vyse, E. R. (1982). Comparative Genetics of Three Trumpeter Swan Populations. Auk, 99(1), 103-108.

Beck PSA, Juday GP, Alix C, Barber VA, Winslow SE, Sousa EE, Heiser P, Herriges JD, Goetz SJ. (2011) Changes in forest productivity across Alaska consistent with biome shift. Ecology Letters, 14(4), 373–379.

Butkauskas, D., Švažas, S., Tubelytė, V., Morkūnas, J., Sruoga, A., Boiko, D. et al (2012). Coexistence and population genetic structure of the whooper swan Cygnus cygnus and mute swan Cygnus olor in Lithuania and Latvia. Central European Journal of Biology, 7(5), 886-894.

Carling, M. D., Brumfield, R. T., and Webster, M (2008). Haldane's Rule in an Avian System: Using Cline Theory and Divergence Population Genetics to Test for Differential Introgression of Mitochondrial, Autosomal, and Sex-Linked Loci Across the Passerina Bunting Hybrid Zone. Evolution, 62(10), 2600-2615.

Chan, C. X., and Ragan, M. (2013). Next-generation phylogenomics. Biology Direct, 8(1), 3.

Chen, X. J. (2013). Mechanism of homologous recombination and implications for aging-related deletions in mitochondrial DNA. Microbiol Mol Biol Rev, 77(3), 476-496.

Dabrowski, A., Fraser, R., Confer, J. L., and Lovette, I. J. (2005). Geographic variability in mitochondrial introgression among hybridizing

populations of Golden-winged (Vermivora chrysoptera) and Blue-winged (V.opinus) Warblers. Conservation Genetics, 6(5), 843-853.

Dingess, E. (2008). The Analysis of Trumpeter and Tundra Swan Hybridization using Sex Chromosome Markers. Unpublished raw data.

Earnst, SL (1994) Tundra Swan habitat preferences during migration in North Dakota. Journal of Wildlife Management, 58(3), 546–551.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 32(5), 1792-1797.

Engelhardt, K. A. M., Kadlec, J. A., Roy, V. L., and Powell, J. A. (2000). Evaluation of translocation criteria: case study with trumpeter swans (Cygnus buccinator). Biological Conservation, 94(2), 173-181.

Evans, ME, William, Sladen WJL (1980) A comparative analysis of the bill markings of whistling and Bewick's swans and out-of-range occurrences of the two taxa. The Auk, 97, 697–703.

Falush D, Stephens M, Pritchard JK (2003) Inference of population structure: Extensions to linked loci and correlated allele frequencies. Genetics, 164, 1567– 1587.

Faivre, B., Secondi, J., Ferry, C., Chastragnat, L., and Cezilly, F. (1999). Morphological variation and the recent evolution of wing length in the Icterine Warbler: a case of unidirectional introgression? Journal of Avian Biology, 30(2), 152.

Grant, P. R., and Grant, B. R. (1992). Hybridization of Bird Species. Science, 256(5054), 193-197.

Hackett, S. J., Kimball, R. T., Reddy, S., Bowie, R. C., Braun, E. L., Braun, M. J. et al (2008) A phylogenomic study of birds reveals their evolutionary history. Science, 320(5884), 1763-1768.

Hansen HA, Shepherd PEK, King JG, Troyer WA (1971) The Trumpeter Swan in Alaska. Wildlife Monographs, 26, 3–83.

Hughes, A. R., and Stachowicz, J. J. (2004). Genetic diversity enhances the resistance of a seagrass ecosystem to disturbance. Proceedings of the National Academy of Sciences, USA, 101(24), 8998-9002.

John, Judy St, Ransler, Findley A., Quinn, Thomas W., and Oyler-Mccance, Sara J. (2006). PRIMER NOTE: Characterization of microsatellite loci isolated in trumpeter swan (Cygnus buccinator). Molecular Ecology Notes, 6(4), 1083-1085.

Johnsgard PA (1974) The taxonomy of the northern swans. Wildfowl, 25, 155 –161.

Johnsgard PA (1978) Revised 2010. University of Nebraska Press. Lincoln, NE.

Lander, E. S., and Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics, 2(3), 231-239.

Lee, J. C., Tsai, L. C., Hwa, P. Y., Chan, C. L., Huang, A., Chin, S. C., et al (2010). A novel strategy for avian species and gender identification using the CHD gene. Mol Cell Probes, 24(1), 27-31.

Limpert, R. J., Sladen, W. J. L., and Allen, J. H. A. (2013). Winter distribution of Tundra Swans Cygnus columbianus breeding in Alaska and Western Canadian Arctic. Wildfowl; 1991: Wildfowl Special Issue No. 1.

Limpert, R. J., Sladen W. J. L., and Hubert A. A. JR. (1991) Winter distribution of Tundra Swans *Cygnus columbianus columbianus* breeding in Alaska and Western Canadian Arctic.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., and Brumfield, R. T. (2013). Applications of next-generation sequencing to phylogeography and phylogenetics. Mol Phylogenet Evol, 66(2), 526-538.

Meng A, Parkin DT (1993) Genetic differentiation in natural populations of swans revealed by DNA fingerprinting. Acta Zoologica Sinica, 39, 209–216.

Metzker, M. L. (2010). Sequencing technologies [mdash] the next generation. Nat Rev Genet, 11(1), 31-46.

Oyler-McCance, S. J., Ransler, F. A., Berkman, L. K., and Quinn, T. W. (2007). A rangewide population genetic study of trumpeter swans. Conservation Genetics, 8(6), 1339-1353.

Panov, E. N., and Pavlova, E. Y. (2010). Observations on a hybrid between the whooper swan (Cygnus cygnus) and the mute swan (Cygnus olor). Biology Bulletin, 37(8), 759-767.

Petit, R. J., and Excoffier, L. (2009). Gene Flow and Species Delimitation. Trends in Ecology and Evolution, 24(7), 386-393.

Pons, J. M., Sonsthagen, S., Dove, C., and Crochet, P. A. (2014). Extensive mitochondrial introgression in North American Great Black-backed Gulls (Larus marinus) from the American Herring Gull (Larus smithsonianus) with little nuclear DNA impact. Heredity (Edinb), 112(3), 226-239.

Portenko, L. A. (1972). The birds of Chukotski Peninsula and Wrangel Island. Vol. 1. Moscow, Nauka Press.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics, 155, 945–959.

Schmidt, J. H., Lindberg, M. S., Johnson, D. S., and Schmutz, J. A. (2009). Environmental and Human Influences on the Trumpeter Swan Occupancy in Alaska. The Condor, 111(2) 266-275.

Secondi, J., Faivre, B., and Bensch, S. (2006). Spreading introgression in the wake of a moving contact zone. Mol Ecol, 15(9), 2463-2475.

Serie, J. R., and Bartonek, J. C. (2013). Population status and productivity of Tundra Swans Cygnus columbianus in North America. Wildfowl; 1991: Wildfowl Special Issue No. 1.

Shea RE, Nelson HK, Gillette LN, King JG, Weaver DK (2002) Restoration of Trumpeter Swans in North America: a century of progress and challenges. Waterbirds, 25 (Special Publication 1), 296–300.

Stein, A. C., Uy, J. A. C., Nürnberger, B. (2006). Unidirectional Introgression Of A Sexually Selected Trait Across An Avian Hybrid Zone: A Role For Female Choice? Evolution, 60(7), 1476-1485.

Travsky A, Beauvais GP (2004) Species assessment for the Trumpeter Swan (Cygnus buccinator) in Wyoming. BLM, Cheyenne, WY.

Treangen, T. J., Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet, 13(1), 36-46.

Wilk, RJ (1988) Distribution, abundance, population structure and productivity of Tundra Swans in Bristol Bay, Alaska. Arctic, 41(4), 288–292.

Wilson, L (2013) Differentiation of the Tundra (*Cygnus columbianus columbianus*) and Trumpeter (*Cygnus buccinators*) Swans and their Hybrids using Microsatellite Regions. (Masters Thesis). Received from Mason Archival Repository Service. (http://hdl.handle.net/1920/8500)

CURRICULUM VITAE

Christopher Kenneth Yesmont graduated from Huntington High School, Huntington, New York, in 2002. He received his Bachelor of Science from the State University of New York, Stony Brook in 2007. He was employed as a Hazardous Materials Specialist for New York City, Environmental Protection for five years, after which he returned to school in 2012 and received his Master of Science in Biology from George Mason University in 2014.