12-2015

A SURVEY OF THE COMMON LOON (Gavia immer) GENOME REVEALS PATTERNS OF NATURAL SELECTION

Zach G. Gayk
zachgayk@gmail.com

A SURVEY OF THE COMMON LOON (*Gavia immer*) GENOME REVEALS PATTERNS OF NATURAL SELECTION

By

Zachary G. Gayk

THESIS

Submitted to
Northern Michigan University
In partial fulfillment of the requirements
For the degree of

MASTER OF SCIENCE

Office of Graduate Education and Research

December 2015

SIGNATURE APPROVAL FORM

Title of Thesis: A SURVEY OF THE COMMON LOON (*Gavia immer*) GENOME REVEALS PATTERNS OF NATURAL SELECTION

This thesis by Zach Gayk is recommended for approval by the student's Thesis Committee and Department Head in the Department of Biology and by the Assistant Provost of Graduate Education and Research.

Committee Chair: Dr. Alec Lindsay Date

First Reader: Dr. Kate Teeter Date

Second Reader (if required): Dr. Kurt Galbreath Date

Third Reader (if required): Dr. Jeff Horn Date

Department Head: Dr. John Rebers Date

Dr. Brian D. Cherry Date
Assistant Provost of Graduate Education and Research

ABSTRACT

A SURVEY OF THE COMMON LOON (*Gavia immer*) GENOME REVEALS PATTERNS OF NATURAL SELECTION

By

Zachary G. Gayk

Comparative genomics has become a viable method for studying the adaptation of species to their environment at the genome level. I investigated this in common loons (*Gavia immer*) by finding signatures of positive selection as evidence for genomic adaptation.

I used Illumina short read sequencing data from a female common loon to produce a fragmented assembly of the common loon (*Gavia immer*) genome. The assembly had a contig N50 of 814 bp, and a total length of 767,326,331 bp. I identified fragments of 13,821 common loon genes and another 348 coding sequences of unknown function, for a total of 14,169 common loon genes. Based on estimates from well-resolved avian genomes, this figure represents 80.7% of common loon genes. I calculated dN/dS ratios between common loon and chicken (*Gallus gallus*) for a high confidence set of 10,106 gene fragments to find genes under positive selection. I found 490 positively selected genes in the common loon that were enriched for a number of protein classes, including those involved in muscle tissue development, immunoglobulin function, hemoglobin iron binding, G-protein receptors, and ATP metabolism.

The signature of positive selection in these areas suggests common loons may have adapted for underwater diving by (1) compensations of the cardiovascular system and oxygen respiration, (2) low-light visual acuity, (3) and improved metabolism. This work represents the first effort to understand the genomic adaptations of the common loon and may have implications for scholars seeking to find genes of interest for population genetic, ecological or conservation studies of the common loon.

ACKNOWLEDGEMENTS

Due to the scale of this project, this thesis work would not have been possible, but for the support and expertise of scholars from multiple disciplines. Special thanks to the members of my committee: (1) Dr. Alec Lindsay for granting me the chance to pursue this work without having any background or prior skill in genomics and being an inspiring science mentor throughout my undergraduate and graduate careers, (2) Dr. Jeff Horn for invaluable assistance with advanced software installation, systems administration, and enthusiastic use of his Williac cluster, (3) Dr. Kate Teeter for bioinformatics expertise, and (4) Dr. Kurt Galbreath for patience with teaching me phylogenetics. In addition, the skill of computer science undergrads Nolan Earl and Justin Syria made them vital assistants in the confusing world of Python scripting. Thanks also for the support of my parents Bill and Sandra Gayk, Diane Raven and Michael Scheiwe, Scott Hickman and Deb Laurie, and fellow Biology graduate students and faculty, especially Andrew Rankin, Cody Norton, Sarah Gallagher, Heather Toman, and Laurel Hill.

This thesis is dedicated to all the staff bioinformaticians at the elite Gavia Genome Group, with the hope that comparative genomics allows us to unravel increasingly tractable problems on the evolution of diversity.

This thesis follows the format of Evolution, to which I intend to submit portions of this manuscript.

TABLE OF CONTENTS

List of Tables v

List of Figures vi

Introduction 1

Background 1
Project Components 3

Methods 4

De Novo Genome Assembly 4

Reference-guided Assembly 5
Genome Annotation 6
Evolutionary Analyses 8

Results 10

Discussion 22

Genome Assembly Quality 22
Genome Size and GC content 23
Gene Identification 24
Evolutionary Analyses and Biological Significance of Genes under Positive Selection 26

Conclusions 33

Literature Cited 34

Appendices 45

Appendix 1. Supplementary Methods 45
Appendix 2. Description of Computer Resources 49
Appendix 3. Location of Gene Data 50

LIST OF TABLES

Table 1. Illumina platform read statistics for the common loon (*Gavia immer*), including Quality Control scores for assignment of correct base call during sequencing. Original versus filtered results indicate the reduction in usable read data post filtering of low-quality reads....................11

Table 2. Assembly contiguity statistics for eight different draft assemblies of the common loon (*Gavia immer*) genome produced with different *k*-mer sizes. The assembly with the maximum N50, or length at which 50 % of sequences met this threshold, is labeled in red..........................12

Table 3. Descriptive statistics for the common loon (*Gavia immer*) genome assembly with highest contig length (*k* = 30) 12

Table 4. CEGMA output displaying percentage of 248 ultra-conserved genes recovered in the k=30 common loon assembly 13

Table 5. Frequency values for each nucleotide in the highest quality (*k* = 30) common loon assembly, including frequency of ambiguous base N and percent GC content.............................14

Table 6. Gene ontology categories for a subset of 5,163 common loon (*Gavia immer*) genes annotated from an assembly using a k-mer size of 30. Counts indicate the total number of genes while Fractions indicate the percent of genes grouping with that ontology term..........................16

Table 7. Positively selected genes in the common loon (*Gavia immer*) that are associated with G-protein receptors and signal transduction 28

LIST OF FIGURES

Figure 1. Quality scores for assignment of correct base call during Illumina sequencing of the common loon (*Gavia immer*), summed across base pair positions from 1-101 in all generated short reads 10

Figure 2. Percent GC content of the common loon k =30 genome assembly. Variation in the percentage of bases comprised of paired G (guanine) and C (cytosine) is shown across assembled bases from 1 to 767, 326, 331 base pairs 14

Figure 3. A semantic grouping of gene ontology clusters of 11,760 common loon genes analyzed with gene ontology enrichment. Each circle represents a distinct gene ontology cluster enriched in the common loon, while circle size indicates the relative proportion of common loon genes enriched for each ontology cluster. Coloration indicates the log *P*-value for confidence that genes enriched are non-randomly associated with a particular ontology term. Spacing of ontology clusters is based on a semantic algorithm that groups ontology clusters according to similar biological function 15

Figure 4. The frequency of pairwise dN/dS values across 10,106 orthologs shared between common loon (*Gavia immer*) and chicken (*Gallus gallus*). Values for dN/dS are interpreted as evidence of purifying selection (dN/dS < 1), neutrality (dN/dS = 0), or positive selection (dN/dS > 1) 18

Figure 5. Frequency of alignment lengths across 10,106 protein coding gene fragments aligned between common loon (*Gavia immer*) and (*Gallus gallus*)....................18

Figure 6. Frequency of alignment lengths across 490 positively selected gene fragments in a pairwise analysis of dN/dS ratios between common loon (*Gavia immer*) and (*Gallus gallus*).....18

Figure 7. Frequency of 490 positively selected genes in a pairwise analysis of dN/dS ratios between common loon (*Gavia immer*) and chicken (*Gallus gallus*) on distinct chromosomes. Chicken chromosomes were used as a reference for common loon chromosomes, with the assumption of 1:1 synteny 19

Figure 8. Kyoto genes and genomes (KEGG) pathway map for common loon genes (*Gavia immer*) ACTC1 and MYH7 involved in cardiac function. Positively selected genes for this pathway in the common loon are shown in red 20

Figure 9. Kyoto genes and genomes (KEGG) pathway map for common loon genes (*Gavia immer*) GNB1 involved in rhodopsin signal transduction. Positively selected genes for this pathway in the common loon are shown in red 21

Figure 10. Kyoto genes and genomes (KEGG) pathway map for common loon genes (*Gavia immer*) B2M and CD8 involved in imunoglobulin function. Positively selected genes for this pathway in the common loon are shown in red 21

INTRODUCTION

Background.—Genomes offer valuable records of the evolutionary forces that have shaped adaptation throughout a species' history. The traditional approach used to study adaptation was to indirectly categorize phenotypic traits (e. g. bill morphology) and then estimate the strength of selection presumably acting on these traits (Schluter and Price 1985) This method has at least two limitations: (1) it can only be applied to species with populations small enough for researchers to track, and (2) measurable morphological traits do not always correspond to genes under selection. For these reasons, the study of adaptation has largely been theoretical (Schluter or limited to well characterized model taxa (Schluter et al. 2008). Genomics has offered an alternative means of studying adaptation and led a reevaluation of traditional studies (Abzhanov et al. 2004; Schluter and Price 2008; Lamichhaney et al. 2015) by providing a method to identify the coding sequences of large numbers of genes—the targets of selection—that is applicable to nearly all organisms as long as DNA can be obtained from them.

Once a large genomic dataset is obtained, studying adaptation for the species of interest typically relies on comparisons to known genomes. Because gene function is highly conserved across even distantly related taxa (Xu et al. 2006; Weber et al. 2014), homologous gene copies can be compared between species to determine if sequence changes have been influenced by natural selection or are instead the result of neutral evolution. Homologous gene copies, called orthologs if they diverged via a speciation event, code for specific amino acids conserved because they have selective consequences for the resulting protein encoded by each gene (Ge et al. 2008). Sequence changes resulting from random mutations that change the amino acid product—called nonsynonymous mutations—should theoretically be removed from populations by purifying selection if these mutations radically alter the protein produced by each gene (Yang and Nielsen 2000). In rare cases, however, the modified protein structure confers an adaptive advantage and such genes are said to be positively selected (Nielsen et al. 2005). Adaptation can be studied via genomic methods by first identifying positively selected genes, and then interpreting the evolutionary significance of their changes. However, genomic studies of positive selection are dependent on acquiring enough genes for a large analysis; this requires a technologically efficient way to sequence and assemble huge amounts of genomic data (Yandell and Ence 2012).

The technology associated with genome sequencing has rapidly progressed since the 1970’s, when the first sequencing techniques using chain-terminating dideoxynucleotides (Sanger et al. 1977) were developed. Using long DNA fragment reads longer than 500 base pairs, Sanger sequencing provided the first reliable method to sequence whole genomes. However, the long read length and run times made this process a time consuming way to sequence large eukaryotic genomes (Schuster 2008). Next-generation sequencing (NGS), developed in 2005, has greatly impacted the field of genomic research by decreasing the required time, number of personnel, and cost of sequencing whole or partial genomes (Schuster 2008). This technological innovation, which relies on the simultaneous sequencing of millions of short DNA fragments, has led to rapid investigation of the structure and function of genomes from taxa across the phylogenetic tree of life (Baker 2012).

The first steps in acquiring a new genome include the assembly of millions of DNA fragments that must be assembled in order to generate a high-quality genome. With approximately 20 different assembly programs currently available (Zhang et al. 2011), the next step in making a genome assembly is choosing the most appropriate assembler program for a particular data set. A number of assemblers have been evaluated for different types of genomic data (Zhang et al. 2011), including eukaryotic versus prokaryotic genomes.

While most assembled vertebrate genomes which have been used in comparative analyses are mammalian (Doyle et al. 2014a), the birds (order: *Aves*) currently lag behind, despite having approximately double the number of extant species (Jetz et al. 2012). To understand the development of key evolutionary trait changes in birds, and how they are encoded at the sequence level, avian genomes need to be sequenced and assembled from both the early and recent taxonomic radiations. Loons (family: *Gaviidae*) represent appropriate avian candidates for such genome studies since extensive biological and conservation research make interpretation of functional genes within an evolutionary context a tractable goal.

I used bioinformatics tools to assemble, and perform comparative genomic analyses of protein coding sequences within the genome of the common loon (*Gavia immer*). Because the common loon genome had not been previously assembled or annotated, this research represented an opportunity to explore the unknown structure of the common loon genome and contribute to our understanding of genome evolution in birds.

The main goal of this project was to create a draft genome assembly of high-quality protein coding fragments of the common loon that could then be annotated and compared to the available orthologs in the chicken (*Gallus gallus*) genome assembly (Hubbard et al. 2002). I used this approach to identify coding sequence changes indicative of positive selection and their potential evolutionary significance.

Project Components and Predictions.—This work was divided into three components: Genome Assembly, Gene Identification, and Evolutionary Analyses. Due to the exploratory and descriptive nature of this work, no specific *a priori* hypotheses or predictions about the function of positively selected genes in common loons were made during the initial assembly phase of this project. Limited work from penguins (Li et al. 2014) suggested that these ocean diving birds have been subject to positive selection in adipocyte, feather keratin, wing development, and opsin genes. Due to the similar ecological foraging characteristics resulting in streamlined, aquatic-focused morphologies, I anticipated that convergent selection pressures might have acted on both loons and penguins. However, no previous studies have elucidated the adaptive significance of gene evolution in freshwater aquatic birds such as loons. In light of this, I interpreted results from selection analyses in the context of common loon specific ecology and behavior.

METHODS

De Novo Assembly.— I used Illumina short read 2 X 100 base pair data generated by Axeq Technologies Inc. (Axeq Technologies report 2011) from a single female common loon for a de novo assembly of the common loon genome. Axeq Technologies' final sequencing output resulted in 499,620,770 sequence reads, comprising 50,461,697,770 total bases in all summed reads (Axeq Technologies report 2011). An initial assembly was included in the Axeq Technologies genome analysis using the program SOAPdenovo (Luo et al. 2012), a short-read assembler developed by the Beijing Genomics Institute (BGI) which uses the de Bruijn graph-based method (Luo et al. 2012); Appendix 1). This cursory SOAP-based assembly was not used in this project as recent improvements in assembly algorithms (Birol et al. 2013) make other assembly programs attractive for bioinformatics analyses (Supplemental Methods 1). For this reason, I repeated the entire assembly process using the raw short read sequence read data from

Axeq Technologies with the ABySS assembler program (Simpson et al. 2009), which performs particularly well with complex vertebrate genomes (Zhang et al. 2011).

I used the message passing interface (MPI) version of ABySS 1.5.2 (Simpson et al. 2009) on a Rocks compute cluster (Appendix 2) to assemble eight versions of the common loon genome, each with different *k*-mer values (Appendix 1), and then compared these assemblies to both the SOAP denovo assembly (Luo et al. 2012) and to each other to assess the resulting quality of each assembly. Metrics used for judging the quality of resulting genome assemblies included contig and scaffold N50—also known as median contig length, percent genome gaps, and percent genome coverage (DOLEŽEL and Bartoš 2005). I attempted to choose the assembly that optimized contig and scaffold N50 based on gene size to achieve the most informative assembly for annotation analyses (Yandell and Ence 2012) described under Gene Identification. The highest quality genome assembly, judged by contig N50, was then annotated in an effort to maximize the length of protein coding regions for evolutionary analyses (Yandell and Ence 2012). For the best assembly, I evaluated genome assembly completeness of entire proteins coding regions with the Core Eukaryotic Genome Mapping Approach (CEGMA) (Parra et al. 2007; Doyle et al. 2014a).

Reference-Guided Assembly: To further improve contig lengths, I selected the ABySS assembly with the largest contigs and scaffolds (assessed via N50 length) and aligned each scaffold in the assembly to the publically available red-throated loon (*Gavia stellata*) genome (Zhang et al. 2014). To align scaffolds in the ABySS assembly to the red-throated loon genome I used the Burrows Wheeler Aligner (BWA) package (Li and Durbin 2010) which can be used in parallel to align large eukaryotic genomes. I utilized tools in the BWA (Li and Durbin 2010) to set gap penalties for sequence alignment and use only correctly filtered reads for the construction of consensus scaffolds. After mapping the ABySS contigs to the red-throated loon genome, I again assessed the N50 of reference-mapped contigs and scaffolds and if they had not improved in length compared to the de novo assembly I used a more complicated approach (Wang et al. 2014). This consisted of extracting the paired end NGS sequence read files and aligning the raw sequence reads to the red-throated loon genome. I then merged consensus sequences from the ABySS (Simpson et al. 2009) assembly alignment with the NGS read alignment using the program SamTools (Li et al. 2009). This resulted in extension of scaffolds suitable for Gene Identification and analysis (Wang et al. 2014). I used BBMAP (Bushnell n.d.) to extract assembly statistics and convert the assembly into fasta format. Read depth (D) of the best genome assembly was then estimated using the following equation (Sims et al. 2014):

D = (total number reads x read length)/ assembly length

This statistic was used in calculations of assembly quality and coverage.

Gene Identification.— I used local BLASTn (McGinnis and Madden 2004) to search the resulting common loon genome assembly, scaffold by scaffold, against the latest (Ensembl release 81) update of the chicken (*Gallus gallus*) coding sequence (Hubbard et al. 2002). I generated BLASTn (McGinnis and Madden 2004) results using a custom-formatted script for 12-column tabular output. Gene names were then retrieved for each hit in the BLAST tabular output by using the Ensembl BiomaRt web interface (Smedley et al. 2009). To do this, the entire chicken coding sequence was first set as the base database. Then all BLAST transcript ids from the BLASTn (McGinnis and Madden 2004) output were set as a filter. Finally, all BLASTn (McGinnis and Madden 2004) transcript ids were mapped to Ensembl gene names within this filtered set of chicken coding sequences. Gene names retrieved for chicken were subsequently mapped back to common loon scaffolds using a custom Python script.

To characterize biological and cellular function of the large number of somewhat inscrutable gene name results, I used a series of publicly available tools to summarize these data as enriched Gene Ontology categories present in the common loon genome. First, I used the GO Term Finder (Boyle et al. 2004) to generate Gene ontology categories for an input list of 11,760 homologs shared between common loon and chicken with sequence fragments matching the mean alignment length of 80 bp or greater. This filter of 80 bp was used to restrict analyses of Gene Ontology to sequence fragments with at least 26 codons. I then used REViGO (Supek et al. 2011) to visualize clusters of gene ontology terms via number of included genes and functionality of shared gene ontology terms. REViGO (Supek et al. 2011) output was visualized via an automatically generated *R* (Venables and Smith 2005) script.

To further reduce redundancy of Gene Ontology results, I performed a Functional Annotation Cluster analysis (Wragg et al. 2015) by restricting input to 2,482 genes with genome assembly fragments of at least 300 base pairs. Gene Ontology categories were then grouped into broader "annotation clusters" using the DAVID web interface (Huang et al. 2009a,b). Each annotation cluster displayed multiple gene ontology categories that associated together in a statistically significant manner. Each cluster first grouped individual genes to a Gene Ontology category using the Fisher's Exact test to examine whether genes could have associated with a particular pathway in a random manner, using the total proportion of genes in the entire pathway as a reference. Then, overall Enrichment Scores were calculated using the negative log of mean P-values from Fisher's Exact Test results. Clusters with higher Enrichment Scores were interpreted to mean that grouped gene ontology categories and the genes within them had a statistically significant association with each other. These methods were used to reduce complexity of large amounts of gene data.

Evolutionary Analyses.—I used the resulting annotation to identify genes within the common loon assembly where evolutionary changes occurred since divergence with the chicken (Hubbard et al. 2002). This was done by calculating all non-synonymous and synonymous substitutions between genes identified in the common loon assembly during gene identification, and chicken coding sequence using the program PAML (Yang 2007), by looking for evidence of selection at the codon level.

First, to prepare data for PAML analyses, a number of data transformations were necessary. All BLASTn (McGinnis and Madden 2004) hits described under Annotation were used as input for pairwise analyses and then filtered by alignment length. As multiple BLASTn (McGinnis and Madden 2004) hits per gene were returned due to short assembly scaffolds, only a single hit with the highest alignment length was kept for each gene. Because some identified gene sequences were fragments too short for biologically meaningful analysis of codons in PAML analyses (Yang and Nielsen 2000), only fragments with the mean alignment length of 80 base pairs or longer were kept. This ensured that all gene fragments analyzed had approximately 26 codons to analyze (but see triplet adjustment).

As Illumina sequencing is not selective as to which DNA strand is amplified (Van Nieuwerburgh et al. 2012), 50% of the gene fragments in the loon assembly were on non-coding rather than coding strands. Because only the coding sequence represents the actual gene sequence translated into amino acid and therefore under selection, I identified template strands using the Genbank (Benson et al. 2005) convention for strand designation. Genbank databases retain only the coding strands of each gene and BLAST (McGinnis and Madden 2004) algorithms match to template queries by reverse complimenting them. All common loon gene fragments analyzed for evidence of selection were first converted to coding strands and then placed into an open reading frame using custom Python scripts (Appendices 1 and 2).

Resulting in-frame ortholog pairs from common loon and chicken were subsequently converted into a single multi-sequence phylip-formatted batch file using a custom Python script. To make sure terminal end fragments were entire codons, I trimmed unpaired nucleotides from the phylip file with a Python script that found the nearest set of nucleotides in the alignment divisible by three.

I input phylip-formatted data into the codeml function of PAML (Yang 2007) to calculate pairwise dN/dS ratios (Yang and Nielsen 2000) between each common loon and chicken ortholog in the set of distinct, frame corrected protein coding sequences identified during annotation. I used dN/dS ratios calculated in PAML (Yang 2007)to determine whether positive selection, purifying selection or neutral evolution had occurred between each common loon and chicken ortholog. To compute dN/dS ratios for each ortholog pair, PAML (Yang 2007)first calculated the number of *possible* nonsynonymous and synonymous sites in each sequence using the HKY85 substitution model (Yang and Nielsen 2000). This was done by estimating the probability of substitutions per codon position being translated into new (nonsynonymous) or functionally equivalent (synonymous) amino acids. Then, the actual distribution of nonsynonymous and synonymous sites per codon was calculated in each sequence and divided by the sites possible to determine dN and dS. Finally, dN and dS were divided to calculate the proportion of nonsynonymous substitutions present in relation to the proportion of synonymous substitutions present, or dN/dS. Orthologs between the common loon and chicken were interpreted to have undergone positive selection if dN/dS was greater than one, neutrally evolved if dN/dS equaled one, and purifying selection if dN/dS was less than one.

Finally, I computed Likelihood Ratio tests (LRT) for each dN/dS ratio by running an additional PAML (Yang 2007) simulation of the likelihood of obtaining similar dN/dS ratios under a null assumption of purely neutral evolution. The resulting likelihood scores were used to calculate an LRT for each gene using the following equation: $-2(\ln L_{null} — \ln L_{estimated})$. This LRT statistic was then compared to a chi-square distribution with one degree of freedom and $\alpha=0.05$.

RESULTS

Genome Assembly: The Illumina 2000 runs used 8kb inserts to generate 100 bp paired-end reads. I only used reads that passed quality control filtering (Figure 1; Table 1) with scores between Q20 and Q30 (99.26 % and 95.43 % confidence in correct base call respectively), (Table 1). This reduced the initial number of 499,620,770 reads with 50,461,697,770 individual bases by 58.26%, leaving 291,098,878 usable reads totaling 26,946,081,239 individual bases post filtering (Table 1).

Figure 1. Quality scores for assignment of correct base call during Illumina sequencing of the common loon (*Gavia immer*), summed across base pair positions from 1-101 in all generated short reads.

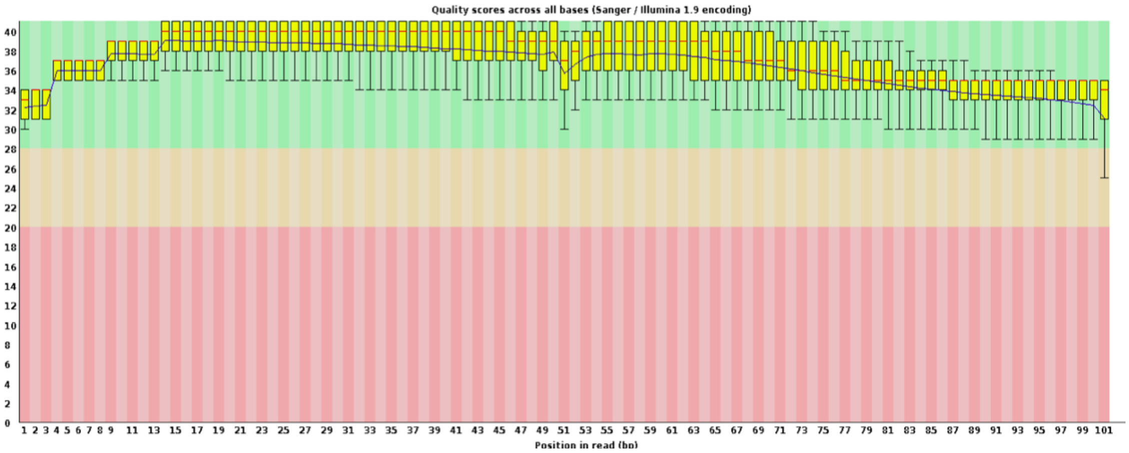

Table 1. Illumina platform read statistics for the common loon (*Gavia immer*), including Quality Control scores for assignment of correct base call during sequencing. Original versus filtered results indicate the reduction in usable read data post filtering of low-quality reads.

	Original		Filtered	
Library	Total Reads	Total Bases	Total Reads	Total Bases
COLO1527	499620770	50,461,697,770	291,098.88	29,946,081,239
Sample	N(%)	GC	Q20	Q30
COLO1527	0.0019	44.53	99.26	95.43

ABySS assemblies of the common loon genome with eight different k-mer sizes ranging from *k*25—*k*64 yielded contig N50 values from 641 to 814 bp in length (Table 2). The *k*-mer size that optimized contig N50 was *k* = 30 with an N50 of 814 bp (Table 2). The *k* = 30 assembly was therefore selected as the best assembly to submit to annotation and evolutionary analyses. Based on BBMAP analyses, the *k* = 30 assembly consisted of 5,237,924 contigs with a total contig length of 767, 326, 331 bp (Table 3). For the *k* = 30 assembly, *k*-mer coverage per ploidy of the diplod sequenced genome was estimated to be approximately 11.84X. Despite the fragmented nature of the genome assembly, 62,409 contigs had lengths of 1 kilobases (kb) or greater. While this comprised only 1.2% of the total contigs and 12.8% of total assembly length, such sequences were of sufficient length for analyses of entire genes and smaller sequences were still suitable for analyses of whole or partial exons. This indicates that 98.8% of the genome assembly consisted of contigs less than 1 kb in length. Similarly, CEGMA output recovered approximately 10% of 248 ultra-conserved proteins (KOG's) summarizing across complete and partial matches (Table 4).

Table 2. Assembly contiguity statistics for eight different draft assemblies of the common loon (*Gavia immer*) genome produced with different *k*-mer sizes. The assembly with the maximum N50, or length at which 50 % of

sequences met this threshold, is labeled in red.

n	n:500	n:N50	min	N80	*N50*	N20	E-size	max	Sum	name	K-mer
4689853	152662	53975	500	579	***761***	1177	935	10767	1.19E+08	unitigs	**37**
4689829	152665	53982	500	579	***761***	1177	935	10767	1.19E+08	contigs	
4689653	152727	53871	500	579	***762***	1178	935	10767	1.19E+08	scaffolds	
2564599	25203	9944	500	542	***639***	867	795	14461	1.70E+07	unitigs	**55**
2564423	25179	9885	500	543	***641***	877	799	14461	1.71E+07	contigs	
2564028	25142	9769	500	543	***645***	902	811	14461	1.72E+07	scaffolds	
5038033	198105	67641	500	591	***802***	1287	1005	7812	1.61E+08	unitigs	**32**
5038000	198106	67653	500	591	***802***	1287	1005	7812	1.61E+08	contigs	
5037795	198153	67499	500	591	***803***	1287	1005	7812	1.62E+08	scaffolds	
3736945	62667	24079	500	554	***678***	955	804	9769	4.42E+07	unitigs	**48**
3736733	62628	24040	500	554	***679***	961	806	9769	4.43E+07	contigs	
3735950	62435	23669	500	555	***684***	986	817	9769	4.45E+07	scaffolds	
1636437	5055	1730	500	542	***655***	1196	1133	12872	3.72E+06	unitigs	**64**
1636380	5054	1717	500	542	***657***	1203	1142	12872	3.83E+06	contigs	
1636124	5088	1698	500	545	***669***	1282	1159	12872	3.83E+06	scaffolds	
6946557	228359	83689	500	578	***747***	1096	873	5000	1.74E+08	unitigs	**25**
6946544	228358	83694	500	578	***747***	1096	873	5000	1.74E+08	contigs	
6946414	228386	83762	500	578	***747***	1096	873	5000	1.74E+08	scaffolds	
5114778	207133	70364	500	593	***809***	1301	1015	7999	1.70E+08	unitigs	**31**
5114751	207137	70181	500	593	***810***	1301	1015	7999	1.70E+08	contigs	
5114566	207200	70239	500	593	***810***	1302	1015	7999	1.70E+08	scaffolds	
5192389	216073	73119	500	595	***814***	1312	1022	7998	1.78E+08	unitigs	**30**
5192361	216073	73130	500	595	***814***	1312	1022	7998	1.78E+08	contigs	
5192194	216128	72984	500	595	***814***	1313	1022	7998	1.78E+08	scaffolds	

Table 3. Descriptive statistics for the common loon (*Gavia immer*) genome assembly with highest contig length (*k* = 30).

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Total Scaffold Length	Total Contig Length	Scaffold Contig Coverage
All	5,237,924	5,238,436	767,438,425	767,326,331	99.99%
50	3,616,441	3,616,953	710,236,525	710,124,431	99.98%
100	2,146,720	2,147,232	604,271,394	604,159,300	99.98%
250	743,885	744,397	394,016,485	393,904,391	99.97%
500	247,247	247,755	223,350,732	223,238,838	99.95%
1 KB	62,044	62,409	98,533,822	98,431,583	99.90%
2.5 KB	5,725	5,731	18,713,830	18,710,728	99.98%
5 KB	231	231	1,310,589	1,310,589	100.00%

Table 4. CEGMA output displaying percentage of 248 ultra-conserved genes recovered in the *k* = 30 common loon assembly.

	#Proteins	% Complete	#Total	Average	%Ortho
Complete	3	1.21	7	2.33	33.33
Group 1	1	1.52	5	5	100
Group 2	0	0	0	0	0
Group 3	1	1.64	1	1	0
Group 4	1	1.54	1	1	0
Partial	23	9.27	52	2.26	65.22
Group 1	4	6.06	10	2.5	75
Group 2	8	14.29	20	2.5	62.5
Group 3	6	9.84	11	1.83	66.67
Group 4	5	7.69	11	2.2	60

From calculations using BBMAP (Li 2015), percent GC content of the *k* = 30 assembly was estimated to be approximately 45.7%, while the proportion comprised of AT content was approximately 54% (Table 5). These numbers differed slightly from the percent GC content recorded in the Axeq assembly report of 44.51% (Axeq unpublished report 2011). Genomic GC content was highly heterogeneous across the genome assembly at a sliding widow size of 10 kb (Figure 2). Local variation in GC content ranged from approximately 30% to 70%. (Figure 2). GC content was especially low within the region spanning 2 Mbp of the genome assembly (Figure 2), which most likely represents a difficult to assemble repetitive region of the common loon genome. CEGMA (Parra et al. 2007) output returned only 10% of 240 possible KOG's. The remaining portions of these analyses focus on the *k* = 30 genome assembly, which is temporarily deposited on the Lindsay Lab iMac under users > zgayk > GAVIABioinformatics, and also on Jeff Horn's WilliacCluster under zgayk > BLASTn or zgayk > K32Assembly (Appendix 3). As further resequencing of the common loon genome is done, this assembly may be useful as a starting assembly for further improvement and eventual submission to Genbank.

Table 5. Frequency values for each nucleotide in the highest quality (*k* = 30) common loon assembly, including frequency of ambiguous base N and percent GC content.

A	C	G	T	N	IUPAC	Other	GC	GC_stdev
0.2714	0.2289	0.2287	0.271	0.0001	0.0001	0	0.4576	0.1076

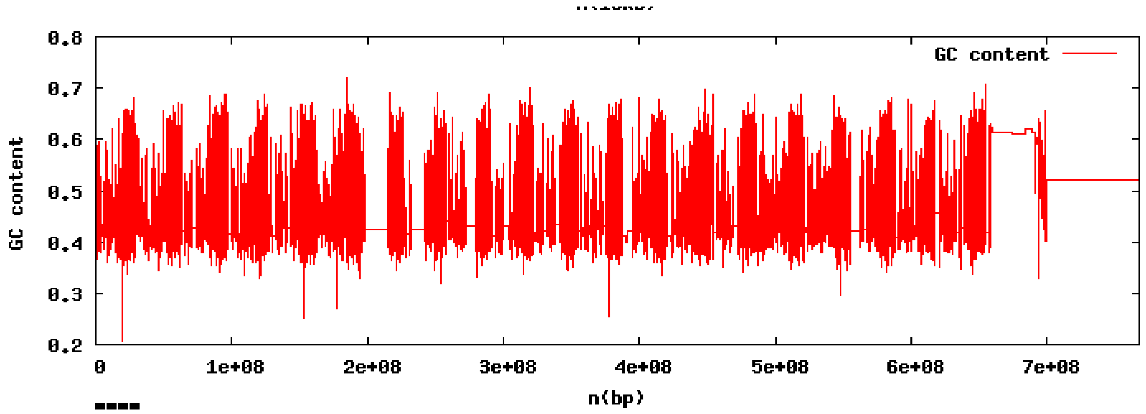

Figure 2. Percent GC content of the common loon *k* =30 genome assembly. Variation in the percentage of bases composed of paired G (guanine) and C (cytosine) is shown across assembled bases from 1 to 767,326,331 base pairs in the total assembly length. Individual values of GC across the genome assembly were plotted using a sliding window analysis set to examine every 10 kilobasepairs (Kbp). Regions with no GC content, shown as white spaces, likely represent sequencing gaps.

Gene Identification: From tabular BLAST output with *k* = 30 assembly scaffolds as query and chicken coding sequence as subject, a database of 136,755 matches to Ensembl transcripts (Hubbard et al. 2002) was returned. Ensembl transcripts had between two to eight result matches to the same scaffold in the *k* = 30 assembly. After filtering with BiomaRt (Smedley et al. 2009), the list of genes identified in the *k* = 30 common loon assembly consisted of 13,821 known chicken genes in Ensembl release 81 and a further 348 unidentified transcripts with unknown function, for a total of 14,169 common loon genes (Supplemental file 1). These results indicate that 80.7% of chicken genes in Ensembl release 81 were identified within the common loon assembly (Supplemental file 1).

From analyses of 11,760 genes with sequence fragments in the common loon assembly at least 80 bp in length, Gene Ontology categories were returned for 1,674 terms (Supplemental File 2). Ontology categories for mitotic cell cycle (54 of 11,760), nuclear division (62 of 11760), and RNA splicing (17 of 11,760) had the highest association with genes from the common loon genome assembly (*P* = 4.80e-310, Supplemental File 2). After submitting a refined list of 2,482 genes that were identified in the common loon genome assembly with fragments ≥ 300 bp, I identified 97 Functional Annotation clusters (Supplemental File 2). Functional Annotation Clusters with the highest enrichment scores (ES) included 21 genes related to nucleotide receptor activity (ES = 11.86, Supplemental File 2), 27 genes related to zinc finger binding (ES = 6.37, Supplemental File 2), and 49 genes related to ion channel activity (ES =4.15, Supplemental File 2). More genes associated with the general domain biological process than any other gene ontology category (Table 6). A grouping of ontology terms by similar biological function and genes per term indicated that there was a clumped distribution of ontology terms, with more genes associating with metabolic process than any other term (Figure 3).

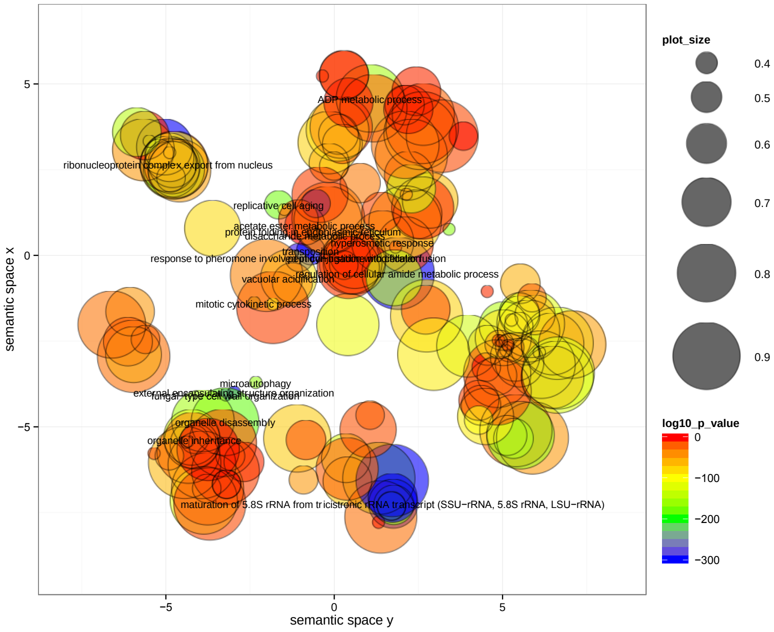

Figure 3. A grouping of gene ontology clusters for 11,760 common loon genes analyzed with gene ontology enrichment. Each circle represents a distinct gene ontology cluster enriched in the common loon, while circle size indicates the relative proportion of common loon genes enriched for each ontology cluster. Coloration indicates the log *P*-value for confidence that genes enriched are non-randomly associated with a particular ontology term. Relative spacing of ontology clusters is based on an algorithm that groups ontology clusters according to similar or overlapping biological function.

Table 6. Gene ontology categories for a subset of 2,482 common loon (*Gavia immer*) genes annotated from an assembly using a k-mer size of 30. Counts indicate the total number of genes while Fractions indicate the percent of genes grouping with that ontology term in relation to a genomic background of 5,163 chicken (*Gallus gallus*) genes used to estimate genome frequency.

GO Class ID	Definitions	Counts	Fractions
GO:0008150	biological_process	1666	32.27%
GO:0008152	metabolism	910	17.63%
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	372	7.21%
GO:0016043	cell organization and biogenesis	329	6.37%
GO:0009058	biosynthesis	263	5.09%
GO:0006996	organelle organization and biogenesis	224	4.34%
GO:0019538	protein metabolism	178	3.45%
GO:0006810	transport	136	2.63%
GO:0007049	cell cycle	135	2.61%
GO:0009056	catabolism	133	2.58%
GO:0006464	protein modification	103	1.99%
GO:0006950	response to stress	87	1.69%
GO:0006259	DNA metabolism	79	1.53%
GO:0007154	cell communication	58	1.12%
GO:0005975	carbohydrate metabolism	56	1.08%
GO:0006629	lipid metabolism	54	1.05%
GO:0000003	reproduction	53	1.03%
GO:0007010	cytoskeleton organization and biogenesis	38	0.74%
GO:0009605	response to external stimulus	34	0.66%
GO:0040007	growth	30	0.58%
GO:0007165	signal transduction	29	0.56%
GO:0006811	ion transport	29	0.56%
GO:0015031	protein transport	27	0.52%
GO:0009628	response to abiotic stimulus	23	0.45%
GO:0006091	generation of precursor metabolites and energy	20	0.39%
GO:0019725	cell homeostasis	18	0.35%
GO:0009653	morphogenesis	17	0.33%
GO:0006412	protein biosynthesis	17	0.33%
GO:0030154	cell differentiation	16	0.31%
GO:0007005	mitochondrion organization and biogenesis	13	0.25%

GO:0040029	regulation of gene expression, epigenetic	12	0.23%
GO:0016049	cell growth	4	0.08%
Total		5163	100.00%

Evolutionary Analyses: Out of 13,211 protein coding sequences, 10,106 were able to be frame-corrected and used in PAML analyses with a high degree of confidence (Supplemental file 3). The remaining 3,105 fragments had no open reading frame despite aligning to known chicken genes. Most sequences had no start codon and multiple stop codons and appeared to be paralogs (sequences resulting from gene duplication events and therefore not under selection).

The distribution of dN/dS ratios across all 10,106 sequences (with alignment lengths between 80—2409 bp) followed a right-skewed distribution, with 5,000 sequences having dN/dS ratios less than one (Figures 4 and 5). For these 5,000 loci, purifying selection was a likely explanation for low dN/dS values. A total of 1018 out of 10,106 had sequences with dN/dS greater than one (10%). However, after removal of sequences with no synonymous mutations, which can lead to spurious dN/dS calculation (Angelis et al. 2014), 700 out of 10,106 gene sequences (6.9%) had dN/dS greater than one. From this gene fragment dataset, likelihood ratio tests resulted in a significant improvement of likelihood scores under the model of positive selection for 490 of 700 genes (70%). This set of 490 genes therefore had statistical support for positive selection between common loon and chicken orthologs and were considered the final set of positively selected genes. Within this set of positively selected genes, alignment sizes varied from 80 to 2409 base pairs in length with a mean alignment length of 199 base pairs (Figure 6). Thirty-six percent of positively selected genes were restricted to chromosomes one and two on the chicken genome, with genes distributed across 27 total chromosomes (Figure 7).

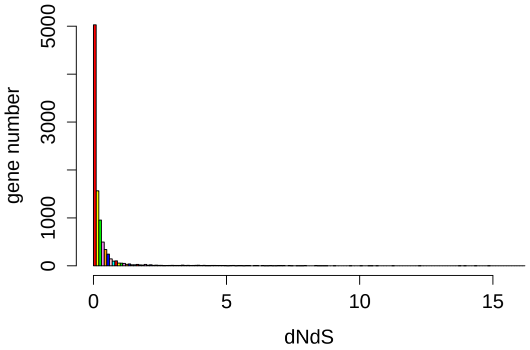

Figure 4. The frequency of pairwise dN/dS values across 10,106 orthologs shared between common loon (*Gavia immer*) and chicken (*Gallus gallus*). Values for dN/dS are interpreted as evidence of purifying selection (dN/dS < 1), neutrality (dN/dS = 1), or positive selection (dN/dS > 1).

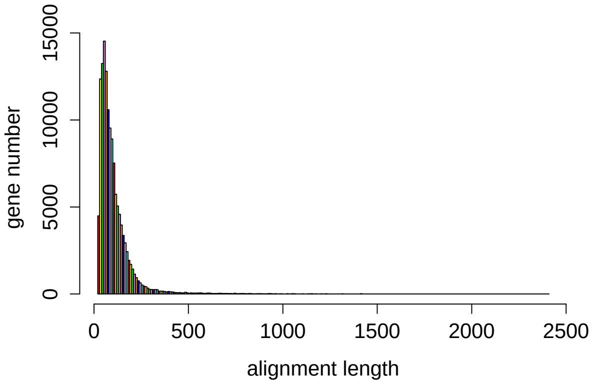

Figure 5. Frequency of alignment lengths across 10,106 protein coding gene fragments aligned between common loon (*Gavia immer*) and (*Gallus gallus*).

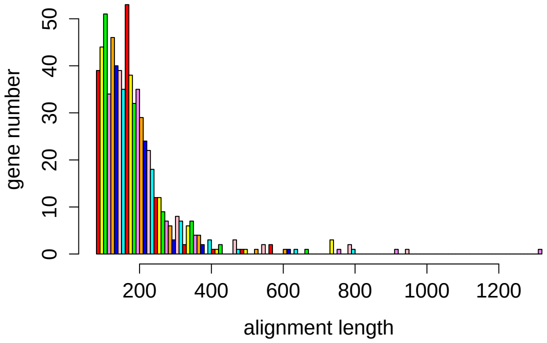

Figure 6. Frequency of alignment lengths across 490 positively selected gene fragments in a pairwise analysis of dN/dS ratios between common loon (*Gavia immer*) and (*Gallus gallus*).

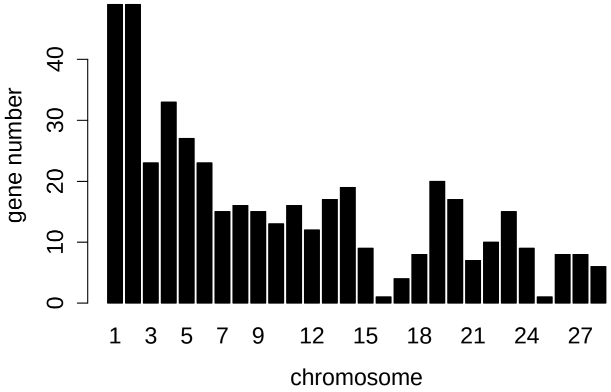

Figure 7. Frequency of 490 positively selected genes in a pairwise analysis of dN/dS ratios between common loon (*Gavia immer*) and chicken (*Gallus gallus*) on distinct chromosomes. Chicken chromosomes were used as a reference for common loon chromosomes, with the assumption of 1:1 synteny.

A functional annotation analysis of the 490 positively selected common loon and chicken orthologs with DAVID (Huang et al. 2009b) revealed 31 enrichment clusters (Supplemental File 3). Enrichment clusters included terms for DNA metabolic process (GO:0006259; ES = 1.470800766912138), muscle tissue development (GO:0060537; ES = 1.2331077295554835), peptidase activity (GO:0070011; ES = 1.0273168307754916), immunoglobulin function (IPR013783:Immunoglobulin-like fold; ES = 0.9526674798243541), hemoglobin iron binding (GO:0005506; ES = 0.3905104581369136), ATP metabolic process (GO:0016887; ES = 0.37637967111475484) nervous system development (GO:0050767; 0.7684493487047698), regulation of apoptosis (GO:0006915; 0.6927690907661083), biosynthetic process (GO:0016481; 0.7552039490261333), and G-protein coupled receptor pathways (GO:0007186; 0.5167606299770078). A number of clusters had overlapping genes, with the most frequently occurring gene under positive selection that clustered with known gene ontology terms consisting of EYA1, GATA2, SMO, and HES5. This may suggest that these genes have broad or developmental functions in molecular as well as biological processes.

Positively selected genes within these clusters for which the biological function can be readily identified are as follows. Muscle tissue development includes *eyes absent* homolog 1 gene (EYA1), which may have a role in organogenesis and eye development. Actin beta-like 2 (ACBL2) and Myosin heavy chain 7 (MYH7) associated with cardiac muscle function (Figure 8). GNB1 is associated with phototransduction (Figure 9). The immunoglobulin function category includes the Beta-2-microglobin gene (B2M), which is the beta subunit of the Major Histocompatibility (MHC) class I proteins (Figure 10). In the iron-binding category, three hemoglobin genes were positively selected including HBE1, which has a role in pulmonary oxygen transport.

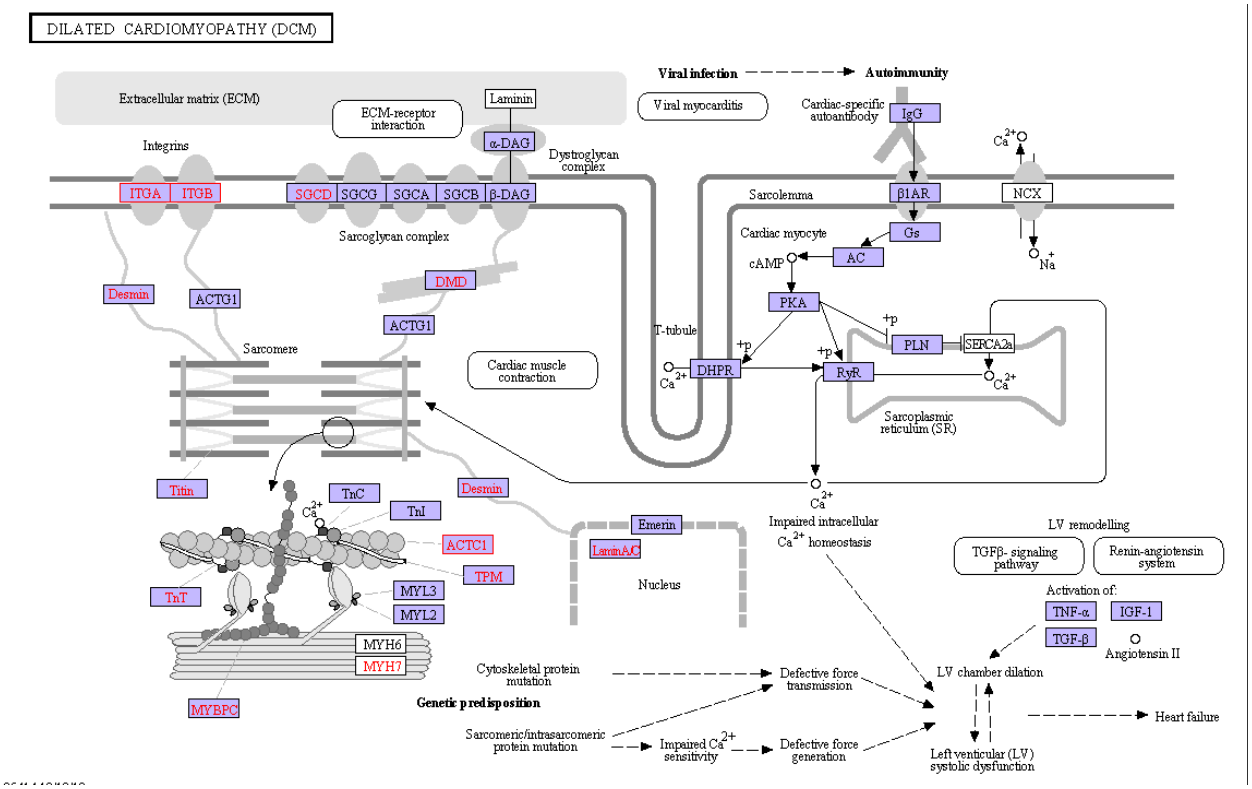

Figure 8. Kyoto genes and genomes (KEGG) pathway map for common loon genes (*Gavia immer*) ACTC1 and MYH7 involved in cardiac function. Positively selected genes for this pathway in the common loon are shown in red. KEGG maps were produced using an online (http://www.genome.jp/kegg/) interface.

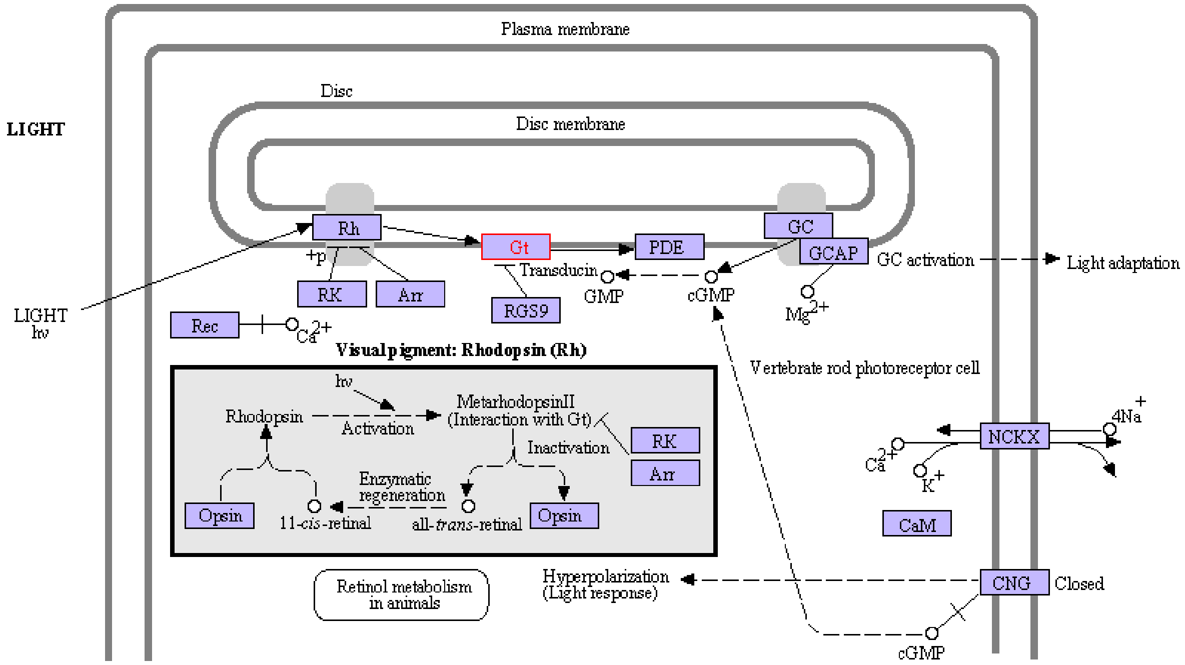

Figure 9. Kyoto genes and genomes (KEGG) pathway map for common loon genes (*Gavia immer*) GNB1 involved in rhodopsin signal transduction. Positively selected genes for this pathway in the common loon are shown in red. KEGG maps were produced using an online (http://www.genome.jp/kegg/) interface.

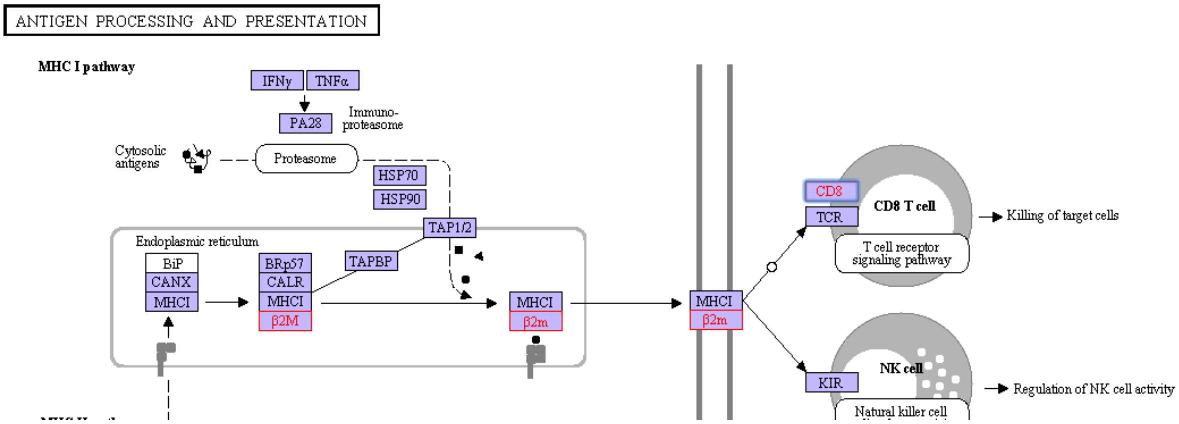

Figure 10. Kyoto genes and genomes (KEGG) pathway map for common loon genes (*Gavia immer*) B2M and CD8 involved in imunoglobulin function. Positively selected genes for this pathway in the common loon are shown in red. KEGG maps were produced using an online (http://www.genome.jp/kegg/) interface.

DISCUSSION

Assembly Quality: The highly fragmented state of the current k = 30 genome assembly, judged by both number of contigs and contig N50, is less than that of highly vetted draft genomes such as chicken, and zebra finch (International Chicken Genome Sequencing Consortium 2014). Contig length appears to be good for 62,044 contigs greater than 1 Kbp, but this comprises a small proportion of total assembly sequence. Recently published genome assemblies from birds including the Hume's ground tit (Cai et al. 2013), golden eagle (Doyle et al. 2014b), and adelie penguin (Zhang et al. 2014) all have contig N50 values in the range of 19-164 Kbp, whereas the current common loon assembly has a contig N50 of 814 bp. Assembly quality most closely approximates that of the black grouse draft assembly (Wang et al. 2014), which had a contig N50 value of 1238 bp. Despite the fragmented nature of the current common loon assembly, contig lengths as indicated by N50 appear to be adequate for identifying protein coding regions of genes (Parra et al. 2009). However, lacking long contiguous scaffolds, the exact placement and order of particular contigs can not be determined and thus is likely inadequate for analyses of synteny (Oleksyk et al. 2012). This is most likely why CEGMA recovered so few ultra-conserved proteins, as this technique is apparently incredibly reliant on long contiguous sequences in its hidden markov predictive model (Parra et al. 2007).

In sum, conclusions about assembly quality must be put in the proper perspective about the resolution desired in analyses. For opportunistic analyses of selection between closely related species making use of available contigs, this assembly has value. However as current goals of other genome assembly studies focused on whole genome structure (rather than genomic studies of adaptation) are to optimize contig N50 for gene size and scaffold N50 to the chromosome level (Yandell and Ence 2012), this common loon assembly will not be accurate for resolving structure of repeat sequences between exons of the same gene. The fragmented nature of the best ABySS assembly can be attributed to several factors including: (1) large (8kb) insert libraries of only one size used in assembly, (2) a short read size of 100 bp, and (3) low sequencing depth. K-mer coverage per ploidy of the diploid loon genome was calculated to be 11.83X, whereas target k-mer coverage should be in the range of 20-30X for a high-quality genome assembly (B. Bushnell pers. com). In this assembly, the odds of correctly assembling each read per ploidy are low given that actual coverage per ploidy is between one-half to one-third the target range for a good genome assembly (B. Bushnell pers. com.), (Yandell and Ence 2012).

The small contig length and the lack of scaffolds is a limitation of the one insert library size (8 kb) providing limited information for resolving genome content above this scale. This is further indicated by the maximal contig length in this assembly nearing but not exceeding 8 kb. This assembly is therefore most appropriate for a targeted identification and analysis of protein coding regions of genes, without concern for producing a complete genome.

Genome Size and GC Content: Avian genomes are between one-third to one-half the size of mammalian genomes. This is generally acknowledged to have evolved as a means of reducing genomic content and weight related to flight (Nam et al. 2010). Estimates from well studied bird genomes suggest typical avian genomes should be in the range of 1—1.5 Gbp with a maximum number of genes under 20,000 (Doyle et al. 2014a). The total consecutive sequence length of the common loon genome assembly was 767 Mpb. Although assembly quality makes estimation of the actual common loon genome size impossible, this figure could indicate we have assembled between 50-76% of the common loon total genomic content using published genome sizes (Doyle et al. 2014a) as a reference.

There have been numerous debates whether GC content, the variation in the structuring of GC bases across genomes, is a result of neutral processes like mutation or drift, or instead selection (Lassalle et al. 2015). In birds and mammals, high GC content is associated with protein coding regions but less so in fish and reptiles, leading some researchers to infer indirect correlations between this process and homeothermy (Galtier et al. 2001). The leading view of higher GC content in protein coding regions (called isochores), however, is the Biased Gene Conversion hypothesis (BGC) (Figuet et al. 2015), which suggests that high GC arises in protein coding regions of some eukaryotes because of a biased mechanism of base repair. During crossing over of homologous chromosomes, the DNA heteroduplex must repair the base composition at sites that are heterozygous. If repair processes are biased toward conversion to GC bases, such GC rich content could accumulate in protein coding regions, provided recombination rates remain high (10^{-8} crossing over per base pair per generation) and effective population sizes approximate 10^{4} individuals (Weissenbach et al. 1992).

Despite the acknowledged issues of contig length, GC content in the common loon assembly was 45.7% overall. Noticeable variation in GC across the common loon genome assembly agrees with the general pattern of isochore development among birds and mammals. Due to the fragmented assembly it is difficult to determine how much common loon GC values are influenced by small contig lengths and gaps in the assembly. That said, a general pattern of at least 20 GC peaks in the 60-70% GC range exist across the assembly; each peak extends for only several hundred thousand kilobases, suggesting these are regions high in protein coding sequences and elevated recombination rates within a matrix of repeats and transposons.

Gene Identification: The 13,821 annotated genes from the common loon assembly (k =

30), comprising matches to 80.7 % of genes in the chicken genome (Ensembl release 81), indicate that our genome project has been successful at identifying a large proportion of avian protein-coding genes despite limited resources. This suggests that the small contig N50 of our assembly was effective for identifying genes widely distributed across the genome, but not resolving entire gene sequences including multiple introns characteristic of eukaryotic genes.

One gene identification technique that limited whole gene retrieval was the use of chicken coding sequences (cds) as a means of identifying loon orthologs. Because chicken cds included the actual DNA sequences of each gene minus introns and untranslated regions (UTR's), but loon genes in the assembly were interspersed with introns and UTRs, BLAST identification of loon cds was limited to homologous regions of contigs until an intron was encountered. Although this was done to limit post-alignment manipulations, a whole genome alignment of chicken and loon genomes may have improved gene lengths slightly. However, I anticipated that gains from a whole genome alignment were not worth the extra processing because of the low N50 value of 814 bp. This is because average gene length (minus introns) in eukaryotes is approximately 1,445 base pairs (Xu et al. 2006), or 631 bases longer than the common loon N50 value. In addition, 90% of common loon and chicken alignments were less than 500 bp in length (Figure 5), with multiple common loon contigs matching the same chicken gene sequence. Resulting annotations are therefore small fragments of loon exons. Entire exons would be needed to provide common loon genes as a harvestable reference set of complete genes for other scholars. One possible use of the current assembly, however is to provide the 13,821 gene fragments so that researchers can pinpoint specific genes of interest for ecological or population genetic studies of common loons and then develop primers around these known gene sequences to obtain the full sequence for such markers.

Evolutionary Analyses and Biological Significance of Genes under Positive Selection: A higher proportion of positively selected genes were clustered on chromosomes one and two than on all other chromosomes. Because the common loon karyotype is not known, but remains highly conserved among birds ((Nam et al. 2010), the use of chicken chromosomes relies on the assumption that common loon and chicken gene level synteny is similar. Clustering of positively selected genes on chromosomes one and two could represent a relative imbalance of positively selected genes per chromosome if the genome assembly was lacking positively selected genes from unassembled chromosomes. This is likely. However, the high number of positively selected genes (PS) on chromosomes one, two, and to a lesser extent five may indicate regions where mutation, and divergence are high. For example, Ellegren et al. (Ellegren et al. 2012) found over 50 divergence peaks in the collared flycatcher (*Ficedula albicolis*) genome irrespective of chromosome size. This would need to be resolved in the common loon with better sequencing, but could indicate that chromosomes one and two contain divergence peaks where selection and linkage disequilibrium and adaptively important genes have evolved at a high rate.

The number of identified positively selected genes across all chromosomes (490), encompassing 3.5% of identified common loon genes, is similar to other comparative genomic studies, although there is some variation with alignment methods and evolutionary distance among genomes compared (Kosiol et al. 2008; Locke et al. 2011). Pairwise analyses used in this study have reduced power to detect positive selection because (1) selection can’t be localized to a particular lineage and (2) a high dN/dS must be obtained across the entire sequence (Yang 2007), which is rare. My use of gene fragments, while reducing comprehensiveness of the genomic survey and likely missing some genes under positive selection, minimizes this second problem. For each positively selected gene in this study, the region where key nonsynonymous mutations occurred is pinpointed to a known alignment length. Future analysis to determine if insertions, deletions or purely missense mutations resulted in these changes would be revealing (Li et al. 2014).

Patterns of gene enrichment suggest that selection since the common loon—chicken split approximately 90 mya (Zhang et al. 2014) has acted on genes related to (1) reorganization of musculature during development (including cardiac morphology), (2) hemoglobin affinity for oxygen, (3) immunoglobulin function related to immune defense, (4) nervous system development and (5) a number of molecular pathways related to DNA metabolic function, peptidase activity, apoptosis, and G-receptor pathways.

A selection analysis of emperor (*Aptenodytes forsteri*) and adelie penguin (*Pygoscelis adeliae*) genomes identified a number of positively selected genes related to Antarctic diving and cold tolerance, and vision in low-light environments (Li et al. 2014). They found: a greater number of β-keratin genes—which comprise 90% of mature feather barbs and barbules—than in any other bird species, a reduction in the number of opsin genes to three trichromatic classes as opposed to four found in most birds as an adaptation to low light environments, positive selection in FASN which encodes lipid metabolism and lipogenesis, and mutation of 17 genes associated with short limb and truncated dorsal morphology for flipper-based diving.

Li et al.'s (Li et al. 2014) focus on identifying genes associated purely with penguin flightlessness and polar marine physiology provide the most phylogenetically similar but still somewhat restricted comparison to loons from an ecological point of view. Although loons (*Gaviiformes*) and penguins (*Sphenisciforme*s) both may have originated in the Southern

Hemisphere (Olson 1992) (the exact origin is unresolved), loons have since evolved to breed and forage on freshwater ponds and lakes during summer and marine environments during winter. Different buoyancy forces and osmotic exchanges exist in freshwater and saltwater environments and loons are one of few migratory aquatic bird classes that exploit both during the same year, or potentially the same day. No loon species have polar distributions and while they do inhabit cold arctic and boreal regions, migration limits the need for the specialized feather and adipose tissue of penguins. I found no positively selected genes among β-keratin or opsin classes of common loon genes. Common loon opsins including OPN3, cOPN5l2, OPNSW and OPN4, all had dN/dS less than 0.02, suggesting high purifying selection constraints on these visual opsins.

All opsins are part of the G-protein receptor superfamily (Provencio 2010), which have direct or indirect roles in signal transduction. Of the four opsins found in this study, OPN3 and OPN4 are not visual opsins but instead have expression in a wider range of tissues and absorption of short-wave visible light used in the development of circadian rhythms (Kosiol et al. 2008). OPN5 and OPNSW are the only visual opsins detected in this study, which absorb blue and UV light peaking in the range of 420 nanometers (nm) wavelength respectively (Provencio 2010). Visual opsins are presumably under strong purifying selection because function contributing specifically to light absorption in both diurnal and crepuscular species is highly conserved (Zhao et al. 2009).

If visual opsin genes controlling direct pigment absorption of light are less likely to be under positive selection, evolutionary changes in avian sight might still occur indirectly through positive selection of G-protein receptors involved in visual pathways related to phototransduction efficiency. Seventeen positively selected common loon genes were enriched for the G-protein receptor pathway (Table 7), which has been shown to be related to sensory perception in species where visual opsins themselves were not under positive selection (Kosiol et al. 2008; Zhao et al. 2009).

Table 7. Positively selected genes in the common loon (*Gavia immer*), which are associated with G-protein receptors and signal transduction.

Gene	Ensembl ID	Align. Length	dN/dS
GPR158	ENSGALT00000012429	303	3.0493
FZD1	ENSGALT00000014752	278	5.7169
TLE4	ENSGALT00000032809	139	1.9604
GNRHR	ENSGALT00000033582	164	2.3876
VIPR1	ENSGALT00000008443	110	2.0182
HCRTR2	ENSGALT00000031641	125	1.8145
CRHR1	ENSGALT00000000503	158	5.8196
SMO	ENSGALT00000040362	212	6.5515
WISP2	ENSGALT00000006604	229	2.2586
GNB1	ENSGALT00000040376	104	5.2283
GALR2	ENSGALT00000002899	784	3.6241
TCTN3	ENSGALT00000017118	81	1.9329
NRG1	ENSGALT00000024876	164	1.4065
IL11RA	ENSGALT00000009397	153	2.4969
NR2F2	ENSGALT00000011332	238	3.8543
DRD3	ENSGALT00000045199	337	1.5363

G-protein receptors (GPR) are important in mediating visual signals by interacting with guanine nucleotide binding proteins. Among common loon genes, GPR158 and GNB1 were grouped with rhodopsin pathways for phototransduction (Figure 8). Although not well known, they may have roles related to G-protein mediation of rhodopsin, the visual opsin in retinal rod cells which has been shown to be related to eyesight in low light conditions (Zhao et al. 2009). EM radiation absorbed by rhodopsin is converted into a signal passed throughout cells by involvement of GPR proteins (Provencio 2010). Rhodopsin has a reddish-purple color and curiously, red retinal color is shared among all loon and nearly all grebes species. This could suggest that underwater vision in deep waters is driven by selection associated with rhodopsin pathways. Rhodopsin is the most photosensitive of all visible opsin pigments and is found in high concentrations among species adapted to light-poor conditions, which often have red eyes. Although speculative, G-protein coupled receptors associated with rhodopsin pathways suggest adaptation in rhodopsin associated genes, possibly for higher concentrations of rhodopsin in loon eyes, or more efficient signal transduction of rhodopsin in low light (Mylvaganam et al. 2006). One possibility is that rhodopsin has been found to photobleach within 20-30 seconds of exposure to light and then has to be replenished in eye tissues quickly for maintenance of sight in low light conditions(Zhao et al. 2009). PS genes in common loons may be associated with a pathway increasing the speed of rhodopsin replenishment (Figure 8), but more work, possibly through RNA-sequencing would be needed to compare timing, duration, and region of expression of GPR158 and GNB1 in loons under and above water.

Another category of enriched PS genes in common loons was the musculature development cluster. Curiously, none of these genes were positively selected in penguins despite similar diving morphology. Eight PS genes (ACTC1, TP63, MYH7, NRG1, NR2F2, SMO, HDAC9, EYA1) were related to development or expression of muscle tissue. Gene EYA1 has a role in regulation of transcription during organogenesis, particularly in ear and eye tissue. SMO has a role in appendage development and muscle tissue homeostasis in mature individuals (Zeng et al. 2015). Taken together, selection in these genes may suggest that substantial reorganizations of the muscular structure of loon have taken place since the split with chicken. At least two potential morphological traits suggest obvious reorganization of musculature and skeletal systems in loons: (1) foot position has been optimized for underwater propulsion such that terrestrial movement is nearly impossible, and (2) at up to 4.2 kg, body mass is extremely heavy for flying birds, and their high wing loading (McIntyre and Olson 1988) requires long takeoffs (up to 200m) and aerobically demanding flight at speeds of 120 km/h [71].

Genes EYA1 and SMO were highly enriched in multiple gene ontology clusters and may have had a role in evolution of posterior appendage and eye tissue. Although it is uncertain, the high statistical support for these genes and high dN/dS (EYA1: 4.318; SMO: 6.5515) indicate these genes have been important in the evolution of the common loon, and that a number of functional changes have occurred in them. Positive selection also has occurred in ACTC1 (actin cardiac muscle) and MYH7 (myosin heavy chain). Both genes encode specific types of actin and myosin involved in muscular contraction of the cardiac muscle. Curiously, mutations in these genes have been shown to cause dilated hypertrophic cardiomyopathy (Jiang et al. 2010; Berti et al. 2015) in humans and mice. A number of studies have reported mutations which appear as dilated cardiomyopathy in humans may actually be a response to hypoxia and cold environments in other species, which evolved convergently among phylogenetically distant high altitude Tibetan ground tits (*Parus humilis*) (Qu et al. 2013), and polar bears (*Ursus arctos*) (Liu et al. 2014). In common loons, positive selection in ACTC1 and MYH7 may be related to oxygen respiration during long underwater dives, which mimics hypoxic conditions. Common loons have been clocked diving for longer than 10 min without surfacing on rare occasions (Nocera and Burgess 2002). The exact adaptations allowing this are unknown, but blood pressure is apparently high among loons (A. Lindsay, pers. comm.), which may increase oxygen saturation of tissues during dives. The NOS3 gene was found to be positively selected in Tibetan antelope, apparently to dilate vessels and increase oxygen saturation in conditions with low atmospheric pO2 (Ge et al. 2013). Exact mechanisms of cardiac evolution in common loons remain unknown, but these results indicate a potential role in adaptation to hypoxia, and possibly for elevated efficiency of heart muscle tissue to increase oxygen saturation during dives.

Further potential evidence for adaptation during long underwater dives may be inferred from the positive selection of three genes associated with hemoglobin affinity for oxygen (HMOX1, HBG1, HBE). HBE and HBG1 are protein-coding forms of hemoglobin epsilon and hemoglobin gamma; normally these genes are only expressed in developing fetal tissues as fetal hemoglobin, although rare conditions allow functional fetal hemoglobins into adulthood (Renneville et al. 2015). Research has suggested that fetal hemoglobins have higher affinities for oxygen than adult hemoglobins (HBA) and maintenance of these genes into adulthood could be related to hypoxic conditions. This however is unlikely without confirmation of expression profiles of adult common loons using RNA sequencing (Alvarez et al. 2015). More likely, selection had acted on functional efficiency of oxygen uptake in developing loon embryos. Taken together with hemoglobin oxygenase (HMOX1), which is thought to have a role in oxygen sensing during hypoxic conditions, positive selection in these three genes suggest that common loons have adapted for oxygen saturation, hypoxia, and diving.

In addition to positive selection in musculature and respiration, I also found signatures of positive selection in genes enriched for immunity and potentially metabolism. Nine genes were positively selected in this category (REL, CD8B, NSFL1C, TGM2, MCAM, NRG1, NTM, B2M, IL11RA), and two (CD8B, B2M) code for microglobulins which are subunits of major histocompatibility class 1 (MHC I) proteins (Alcaide et al. 2014). MHC I are a well studied group of proteins in birds that have been shown to have a role in distinguishing between foreign and parent cells, and also mate selection (Penn and Potts 1999). B2M is the gene for the β-2 microglobulin subunit of the MHC I protein (Taniguchi et al. 2014), while CD8B has a role in antigen recognition and presentation (Figure 10). Further a number of genes were positively selected for regulation of apoptosis, which have been found to be associated with promoting cell death in an immune context.

Common loons also show evidence for positive selection in genes associated with reorganization of neural tissues. This agrees with the massive avian comparative genomics project of Zhang [17], who found rearrangements of spinal and neuromuscular genes were one of the largest groups of enriched genes across 48 phylogenetically representative bird species. In common loons these include NRG1 (neuroregulin), which functions as a signaling molecule important in establishing acetylcholine receptors at neuromuscular junctions, and HES, which is involved in neurogenesis. Specifically, HES is a transcriptional regulator of proteins involved in Notch signaling (Carvalho et al. 2015), which is important for promoting proliferative signaling in neural development.

Several hundred genes were enriched among ontology categories for ATPase activity, binding of ATP, and regulation of transcription. A number of ATPase genes were positively selected in Tibetan antelope, likely as an adaptation for increased aerobic activity and high altitude metabolism (Ge et al. 2013). In common loons, I have already suggested that high aerobic and metabolic costs of a physiology adapted foremost to deep-water diving, but also long distance (trans-continental) aerial migration indicate disparate selection pressures have shaped loon evolution. The optimal morphology for diving, (i.e high mass, easily concealed wings, ventrally positioned feet) presents severe trade-offs for flight, as high mass and narrow wings make it difficult for loons to become airborne and require high flight speeds once aloft (Evers et al. 2010). I hypothesize that a number of genes associated with ATP and metabolism have been positively selected in common loons to maximize energy production in these environments.

Conclusions and Future Directions.—Most of the hypotheses for positive selection in common loons remain speculative unless confirmed through additional study. However, now that candidate PS genes are available, future work could examine the expression of these genes through non-invasive RNA-sequencing (Alvarez et al. 2015); in particular the exact mechanisms through which low-light phototransduction and oxygen saturation have evolved should be elucidated. Expanding the analyses of positive selection to multiple species with differing scales of phylogenetic relatedness (including the red-throated loon (*Gavia stellata*)) would allow selection to be interpreted in a broader evolutionary context. The most compelling approach to interpret the adaptive context of common loon evolution would integrate high-throughput genomic data (as in this study), and established common loon natural history, with direct hypothesis driven tests. It is the latter that is lacking in this study, but in this work I have now provided a reference set of common loon genes other scholars can examine, and perhaps sequence fully for future studies. In addition, this study demonstrates that high throughput genome assembly methods can be used inexpensively to identify coding regions of genes. As NGS sequencing continues to become more common and whole or partial genomic data become available for large numbers of species, more studies may develop tools to harvest incomplete genome assemblies for evolutionary comparative analyses.

LITERATURE CITED

Abzhanov, A., M. Protas, B. R. Grant, P. R. Grant, and C. J. Tabin. 2004. Bmp4 and morphological variation of beaks in Darwin's finches. Science 305:1462–1465. American Association for the Advancement of Science.

Alcaide, M., J. Muñoz, J. Martínez-de la Puente, R. Soriguer, and J. Figuerola. 2014. Extraordinary MHC class II B diversity in a non-passerine, wild bird: the Eurasian Coot *Fulica atra* (Aves:

Rallidae). Ecol Evol 4:688–698.

Alkan, C., S. Sajjadian, and E. E. Eichler. 2011. Limitations of next-generation genome sequence assembly. Nat. Methods 8:61–65.

Alvarez, M., A. W. Schrey, and C. L. Richards. 2015. Ten years of transcriptomics in wild populations: what have we learned about their ecology and evolution? Molecular Ecology 24:710–725.

Angelis, K., M. Dos Reis, and Z. Yang. 2014. Bayesian estimation of nonsynonymous/synonymous rate ratios for pairwise sequence comparisons. Mol. Biol. Evol. 31:1902–1913.

Axeq Technologies. 2011. De novo assembly analysis report. Unpublished report.

Baker, M. 2012. De novo genome assembly: what every biologist should know. Nat. Methods 9:333.

Benson, D. A., I. Karsch-Mizrachi, and D. J. Lipman. 2005. GenBank. Nucleic Acids.

Berti, F., J. M. Nogueira, S. Wöhrle, D. R. Sobreira, K. Hawrot, and S. Dietrich. 2015. Time course and side-by-side analysis of mesodermal, pre-myogenic, myogenic and differentiated cell markers in the chicken model for skeletal muscle formation. J. Anat. 227:361–382.

Birol, I., A. Raymond, S. D. Jackman, S. Pleasance, R. Coope, G. A. Taylor, M. M. S. Yuen, C. I. Keeling, D. Brand, B. P. Vandervalk, H. Kirk, P. Pandoh, R. A. Moore, Y. Zhao, A. J. Mungall, B. Jaquish, A. Yanchuk, C. Ritland, B. Boyle, J. Bousquet, K. Ritland, J. Mackay, J. Bohlmann, and S. J. M. Jones. 2013. Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497.

Boyle, E. I., S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry, and G. Sherlock. 2004.

GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20:3710–3715.

Bushnell, B. n.d. [CITATION][C]. URL http://sourceforge. net/projects/bbmap

Cai, Q., X. Qian, Y. Lang, Y. Luo, S. Pan, Y. Hui, C. Gou, Y. Cai, M. Hao, J. Zhao, S. Wang, Z. Wang, X. Zhang, J. Liu, L. Luo, Y. Li, J. Wang, R. He, F. Lei, and J. Xu. 2013. The genome sequence of the ground tit Pseudopodoces humilis provides insights into its adaptation to high altitude. Genome Biol 14:R29.

Carvalho, F. L. F., L. Marchionni, A. Gupta, B. A. Kummangal, E. M. Schaeffer, A. E. Ross, and D. M. Berman. 2015. HES6 promotes prostate cancer aggressiveness independently of Notch signalling. J. Cell. Mol. Med. 19:1624–1636.

DOLEŽEL, J., and J. Bartoš. 2005. Plant DNA flow cytometry and estimation of nuclear genome size. Annals of Botany.

Doyle, J. M., T. E. Katzner, P. H. Bloom, Y. Ji, B. K. Wijayawardena, and J. A. DeWoody. 2014a. The Genome Sequence of a Widespread Apex Predator, the Golden Eagle (*Aquila chrysaetos*). PLoS ONE 9:e95599.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H. Mäkinen, K. Nadachowska-Brzyska, A. Qvarnström, S. Uebbing, and J. B. W. Wolf. 2012. The genomic landscape of species divergence in Ficedula flycatchers. Nature, doi: 10.1038/nature11584.

Figuet, E., M. Ballenghien, J. Romiguier, and N. Galtier. 2015. Biased gene conversion and GC-content evolution in the coding sequences of reptiles and vertebrates. Genome Biol Evol 7:240–250.

Galtier, N., G. Piganeau, D. Mouchiroud, and L. Duret. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.

Ge, G., L. Cowen, X. Feng, and G. Widmer. 2008. Protein coding gene nucleotide substitution pattern in the apicomplexan protozoa *Cryptosporidium parvum* and *Cryptosporidium hominis*. Comp. Funct. Genomics 879023.

Ge, R.-L., Q. Cai, Y.-Y. Shen, A. San, L. Ma, Y. Zhang, X. Yi, Y. Chen, L. Yang, Y. Huang, R. He, Y. Hui, M. Hao, Y. Li, B. Wang, X. Ou, J. Xu, Y. Zhang, K. Wu, C. Geng, W. Zhou, T. Zhou, D. M. Irwin, Y. Yang, L. Ying, H. Bao, J. Kim, D. M. Larkin, J. Ma, H. A. Lewin, J. Xing, R. N. Platt, D. A. Ray, L. Auvil, B. Capitanu, X. Zhang, G. Zhang, R. W. Murphy, J. Wang, Y.-P. Zhang, and J. Wang. 2013. Draft genome sequence of the Tibetan antelope. Nat Comms 4:1858.

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009a. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Research 37:1–13.

Huang, D. W., B. T. Sherman, and R. A. Lempicki. 2009b. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57.

Hubbard, T., D. Barker, E. Birney, and G. Cameron. 2002. The Ensembl genome database project. Nucleic Acids.

International Chicken Genome Sequencing Consortium. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature, 432*(7018), 695–716. doi:10.1038/nature03154.

Jetz, W., G. H. Thomas, J. B. Joy, K. Hartmann, and A. O. Mooers. 2012. The global diversity of birds in space and time. Nature 491:444–448.

Jiang, H. K., G. R. Qiu, J. Li-Ling, N. Xin, and K. L. Sun. 2010. Reduced ACTC1 expression might play a role in the onset of congenital heart disease by inducing cardiomyocyte apoptosis. Circulation Journal.

Kosiol, C., T. Vinar, R. R. da Fonseca, M. J. Hubisz, C. D. Bustamante, R. Nielsen, and A. Siepel. 2008. Patterns of Positive Selection in Six Mammalian Genomes. PLoS Genet 4:e1000144.

Lamichhaney, S., J. Berglund, M. S. Almén, K. Maqbool, M. Grabherr, A. Martinez-Barrio, M. Promerová, C.-J. Rubin, C. Wang, N. Zamani, B. R. Grant, P. R. Grant, M. T. Webster, and L. Andersson. 2015. Evolution of Darwin/'s finches and their beaks revealed by genome sequencing. Nature 518:371–375. Nature Publishing Group.

Lassalle, F., S. Périan, T. Bataillon, X. Nesme, L. Duret, and V. Daubin. 2015. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet 11:e1004941.

Li, C., Y. Zhang, J. Li, L. Kong, H. Hu, H. Pan, L. Xu, Y. Deng, Q. Li, L. Jin, H. Yu, Y. Chen, B. Liu, L. Yang, S. Liu, Y. Zhang, Y. Lang, J. Xia, W. He, Q. Shi, S. Subramanian, C. D. Millar, S. Meader, C. M. Rands, M. K. Fujita, M. J. Greenwold, T. A. Castoe, D. D. Pollock, W. Gu, K. Nam, H. Ellegren, S. Y. Ho, D. W. Burt, C. P. Ponting, E. D. Jarvis, M. T. P. Gilbert, H. Yang, J. Wang, D. M. Lambert, J. Wang, and G. Zhang. 2014. Two Antarctic penguin genomes reveal insights into their evolutionary history and molecular changes related to the Antarctic environment. 3:1–15.

Li, H. 2015. Correcting Illumina sequencing errors for human data. Arxiv.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 26:589–595.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. 2009. The sequence alignment/map format and SAMtools.

Lindsay, A. R. (2002). Molecular and Vocal Evolution in Loons (Aves: Gaviiformes) Doctoral dissertation, University of Michigan.

Liu, S., E. D. Lorenzen, M. Fumagalli, B. Li, K. Harris, Z. Xiong, L. Zhou, T. S. Korneliussen, M. Somel, C. Babbitt, G. Wray, J. Li, W. He, Z. Wang, W. Fu, X. Xiang, C. C. Morgan, A. Doherty, M. J. O'Connell, J. O. McInerney, E. W. Born, L. Dalén, R. Dietz, L. Orlando, C. Sonne, G. Zhang, R. Nielsen, E. Willerslev, and J. Wang. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. Cell 157:785–794.

Locke, D. P., L. W. Hillier, W. C. Warren, K. C. Worley, L. V. Nazareth, D. M. Muzny, S.-P. Yang, Z. Wang, A. T. Chinwalla, P. Minx, M. Mitreva, L. Cook, K. D. Delehaunty, C. Fronick, H. Schmidt, L. A. Fulton, R. S. Fulton, J. O. Nelson, V. Magrini, C. Pohl, T. A. Graves, C. Markovic, A. Cree, H. H. Dinh, J. Hume, C. L. Kovar, G. R. Fowler, G. Lunter, S. Meader, A. Heger, C. P. Ponting, T. Marques-Bonet, C. Alkan, L. Chen, Z. Cheng, J. M. Kidd, E. E. Eichler, S. White, S. Searle, A. J. Vilella, Y. Chen, P. Flicek, J. Ma, B. Raney, B. Suh, R. Burhans, J. Herrero, D. Haussler, R. Faria, O. Fernando, F. Darré, D. Farré, E. Gazave, M. Oliva, A. Navarro, R. Roberto, O. Capozzi, N. Archidiacono, G. D. Valle, S. Purgato, M. Rocchi, M. K.

Konkel, J. A. Walker, B. Ullmer, M. A. Batzer, A. F. A. Smit, R. Hubley, C. Casola, D. R. Schrider, M. W. Hahn, V. Quesada, X. S. Puente, G. R. Ordoñez, C. López-Otín, T. Vinar, B. Brejova, A. Ratan, R. S. Harris, W. Miller, C. Kosiol, H. A. Lawson, V. Taliwal, A. L. Martins, A. Siepel, A. RoyChoudhury, X. Ma, J. Degenhardt, C. D. Bustamante, R. N. Gutenkunst, T. Mailund, J. Y. Dutheil, A. Hobolth, M. H. Schierup, O. A. Ryder, Y. Yoshinaga, P. J. de Jong, G. M. Weinstock, J. Rogers, E. R. Mardis, R. A. Gibbs, and R. K. Wilson. 2011. Comparative and demographic analysis of orang-utan genomes. Nature 469:529–533.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18.

McGinnis, S., and T. L. Madden. 2004. BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic Acids Research.

McIntyre, J. W., and A. Olson. 1988. The Common Loon: spirit of northern lakes.

Mylvaganam, G. H., T. L. McGee, E. L. Berson, and T. P. Dryja. 2006. A screen for mutations in the transducin gene GNB1 in patients with autosomal dominant retinitis pigmentosa. Mol Vis 12:1496–1498.

Nam, K., C. Mugal, B. Nabholz, H. Schielzeth, J. B. W. Wolf, N. Backström, A. Künstner, C. N. Balakrishnan, A. Heger, C. P. Ponting, D. F. Clayton, and H. Ellegren. 2010. Molecular evolution of genes in avian genomes. Genome Biol 11:R68.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J Sninsky, M. D. Adams, and M. Cargill. 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. Plos Biol 3:e170.

Nocera, J. J., and N. M. Burgess. 2002. Diving schedules of Common Loons in varying environments. Can. J. Zool. 80:1643–1648.

Oleksyk, T. K., J.-F. Pombert, D. Siu, A. Mazo-Vargas, B. Ramos, W. Guiblet, Y. Afanador, C. T. Ruiz-Rodriguez, M. L. Nickerson, D. M. Logue, M. Dean, L. Figueroa, R. Valentin, and J.-C. Martinez-Cruzado. 2012. A locally funded Puerto Rican parrot (Amazona vittata) genome sequencing project increases avian data and advances young researcher education. Gigascience 1:14.

Parra, G., K. Bradnam, and I. Korf. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23:1061–1067.

Parra, G., K. Bradnam, Z. Ning, T. Keane, and I. Korf. 2009. Assessing the gene space in draft genomes. Nucleic Acids Research 37:289–297.

Penn, D. J., and W. K. Potts. 1999. The Evolution of Mating Preferences and Major Histocompatibility Complex Genes. The American Naturalist 153:145–164. The American Society of Naturalists.

Pittman, J. A. 1953. Direct observation of the flight speed of the Common Loon. Wilson Bull. 65:213.

Provencio, I. (2010). Shedding light on photoperiodism. Proceedings of the National Academy of Sciences of the United States of America, *107*(36), 15662–15663. doi:10.1073/pnas.1010370107

Qu, Y., H. Zhao, N. Han, G. Zhou, G. Song, Bin Gao, S. Tian, J. Zhang, R. Zhang, X. Meng, Y. Zhang, Y. Zhang, X. Zhu, W. Wang, D. Lambert, P. G. P. Ericson, S. Subramanian, C. Yeung, H. Zhu, Z. Jiang, R. Li, and F. Lei. 2013. Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. Nat Comms 4:2071. Nature Publishing Group.

Renneville, A., P. Van Galen, M. C. Canver, M. McConkey, J. M. Krill-Burger, D. M. Dorfman, E. B. Holson, B. E. Bernstein, S. H. Orkin, D. E. Bauer, and B. L. Ebert. 2015. EHMT1 and EHMT2 inhibition induces fetal hemoglobin expression. Blood 126:1930–1939.

Sanger, F., S. Nicklen, and A. R. Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A. 74:5463–5467.

Schluter, D., and T. D. Price. 2008. Likelihood of Ancestor States in Adaptive Radiation. Evolution 1–14.

Schluter, D., T. D. Price, and P. R. Grant. 1985. Ecological character displacement in Darwin's finches. Science 227:1056–1059. American Association for the Advancement of Science.

Schuster, S. C. 2008. Next-generation sequencing transforms today's biology. Nat. Methods 5:16–18.

Simpson, J. T., K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones, and I. Birol. 2009. ABySS: a parallel assembler for short read sequence data. Genome Research 19:1117–1123.

Sims, D., I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15:121–132.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). BioMart – biological queries made easy. *BMC Genomics*, *10*(1), 22. doi:10.1186/1471-2164-10-

22.

Supek, F., M. Bošnjak, N. Škunca, and T. Šmuc. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE.

Taniguchi, Y., K. Matsumoto, H. Matsuda, T. Yamada, T. Sugiyama, K. Homma, Y. Kaneko, S. Yamagishi, and H. Iwaisaki. 2014. Structure and Polymorphism of the Major Histocompatibility Complex Class II Region in the Japanese Crested Ibis, Nipponia nippon. PLoS ONE 9:e108506.

Van Nieuwerburgh, F., R. C. Thompson, J. Ledesma, D. Deforce, T. Gaasterland, P. Ordoukhanian, and S. R. Head. 2012. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. Nucleic Acids Research 40:e24–e24.

Venables, W. N., and D. M. Smith. 2014. The R development core team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Wang, B., R. Ekblom, I. Bunikis, H. Siitari, and J. H. glund. 2014. Whole genome sequencing of the black grouse (Tetrao tetrix): reference guided assembly suggests faster-Z and MHC evolution. BMC Genomics 15:1–13. BMC Genomics.

Weber, C. C., B. Nabholz, and J. Romiguier. 2014. Kr/Kc but not d (N)/d (S) correlates positively with body mass in birds, raising implications for inferring lineage-specific selection. Genome Biology.

Weissenbach, J., G. Gyapay, C. Dib, A. Vignal, J. Morissette, P. Millasseau, G. Vaysseix, and M. Lathrop. 1992. A second-generation linkage map of the human genome. Nature 359:794–801.

Wragg, D., A. S. Mason, L. Yu, R. Kuo, R. A. Lawal, T. T. Desta, J. M. Mwacharo, C.-Y. Cho, S. Kemp, D. W. Burt, and O. Hanotte. 2015. Genome-wide analysis reveals the extent of EAV-HP integration in domestic chicken. BMC Genomics 16:784.

Xu, L., H. Chen, X. Hu, R. Zhang, Z. Zhang, and Z. W. Luo. 2006. Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Mol. Biol. Evol. 23:1107–1108.

Yandell, M., and D. Ence. 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet 13:329–342.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.

Young, A. L., H. O. Abaan, D. Zerbino, J. C. Mullikin, E. Birney, and E. H. Margulies. 2010. A new strategy for genome assembly using short sequence reads and reduced representation libraries. Genome Research 20:249–256.

Zeng, N., J. Wu, W.-C. Zhu, B. Shi, and Z.-L. Jia. 2015. Evaluation of the association of polymorphisms in EYA1, environmental factors, and non-syndromic orofacial clefts in Western Han Chinese. J. Oral Pathol. Med., doi: 10.1111/jop.12311.

Zhang, G., B. Li, C. Li, M. T. P. Gilbert, E. D. Jarvis, and J. Wang. 2014. Comparative genomic data of the Avian Phylogenomics Project. 3:1–8.

Zhang, W., J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen. 2011. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. PLoS ONE 6:e17915.

Zhao, H., B. Ru, E. C. Teeling, C. G. Faulkes, S. Zhang, and S. J. Rossiter. 2009. Rhodopsin molecular evolution in mammals inhabiting low light environments. PLoS ONE 4:e8326.

APPENDICES:

Appendix 1. Supplemental Methods and Overview

De Novo Assembly Overview.—Assembly of NGS genomes uses either the de Bruijn graph or overlap-layout consensus algorithms. De Bruijn graph-based assemblers are currently considered more advanced than overlap-layout-consensus assemblers and so are more commonly used for large genomes (Compeau et al. 2011). Genome assembly based on the de Bruijn graph algorithm starts by dividing short NGS sequence reads (roughly 60-150 base pairs) into smaller fragments called *k*-mers of length *k* (Baker 2012). Although *k* should be carefully chosen, the *k*-mer method is used to decrease the computational difficulty of assembling billions of read fragments (Alkan et al. 2011). All *k*-mers with overlap between the DNA sequence are then aligned based on the number of shared nucleotides, and the actual genome sequence can then be determined. In this process, all contiguous overlapping fragment reads are aligned and assembled into the largest possible contiguous sequences without gaps in the underlying nucleotides. Such sequences, called contigs, represent the main result output of a good genome assembly.

After DNA assembly has reached the stage where fragments consist of numerous contigs isolated by remaining gaps where the genome sequence is ambiguous, the process of scaffolding begins. Most assembler programs now incorporate techniques to scaffold DNA into “super-contigs”, or contigs joined to other contigs by gaps in the sequence read quality (Swain et al. 2012). Scaffolds are assembled using paired-end read data, which allows the length comprising the gap sequence to be estimated [32].

Reference Guided Assembly Overview.—I improved the quality of the best assembly produced by ABySS [16] using a reference guided assembly approach. Reference-guided genome assembly can be carried out one of two ways, which I will describe below. The first technique consists of producing a de novo assembly and then aligning these contigs to a reference genome [19]. In the second technique, raw sequence reads generated via NGS are aligned to a reference genome and then the contiguous read fragments are exported and used to produce a genome assembly [19]. In this research, I proposed to use the first technique because, as above, ABySS was used to produce an initial de novo genome assembly. Reference –guided assembly can be particularly useful when there is not great confidence in a de novo assembly and time and energy costs necessary to improve the assembly preclude further sequencing. Particularly if the assembly results in short contigs or scaffolds and there is no time to re-sequence the taxon under study, reference-guided assembly may be useful to improve the genome assembly. A technical concern in this study is that source NGS data for this project was produced using only one read library size (8 kb). A number of read libraries constructed of different sized inserts improve the quality of a de novo assembly because this yields information on where the NGS reads were sequenced from and the likelihood of one fragment adjoining another across gaps in the scaffolding process [65]. Scaffolding will therefore be difficult in any de novo assembly generated from our data with only 8 kb libraries. A reference-guided assembly approach may rectify this problem by building consensus scaffolds based on the proximity of contigs mapped to a reference genome.

Genome Annotation Overview.—Gene annotation is a complex process, requiring many software programs to achieve a meaningful annotation. Because of this, annotation pipelines have now been developed that integrate all software components into single pipelines. The most common approach for genome annotation consists of using basic local alignment search tool algorithms (BLAST) to search genome databases for similar sequences present in the unannotated genome assembly. RNA and expressed sequenced tag (EST) information can also be used in some programs. Before using BLAST searches, several annotation pipelines now incorporate programs to analyze regions of the assembled genome for repeat sequences arising from transposable elements. One such program, Repeat Masker, masks repeat regions in the genome assembly so that BLAST searches do not identify chimeric coding sequences within this region [66]. Subsequently, all potential coding regions are identified using BLAST searches to align with reference genomes. Then, several programs such as SNAP and AUGUSTUS can be used to make *ab initio* gene predictions for remaining regions of the genome that may include exons not otherwise identified using available reference genomes. The advantage of using an annotation pipeline software, such as MAKER2 is that once set to examine a specific genome assembly, the annotation process can be semi-automated. This significantly decreases the time spent otherwise manually annotating genome assemblies. To vouchsafe automated annotation settings and ensure they are giving accurate results, many annotations now include a manual review of aspects of the genome assembly. Proposed annotation methods included using MAKER2 (Holt and Yandell 2011). MAKER2 incorporates four separate programs to build a pipeline for annotation (Doyle et al. 2014) including Repeat-Masker, BLAST, AUGUSTUS and SNAP. First, Repeat Masker was used to identify regions of sequence repeats in the assembly, including those arising from LINE, and SINE repeats. However, problems implementing the MPI (message passing interface) version of MAKER limited our work to only identifying repeat regions of the common loon genome. I had planned to use MAKER2 to initiate BLAST searches based on RNA, and protein data from other available bird genomes. The programs AUGUSTUS and SNAP were to be used to make *ab initio* predictions of gene function for remaining unidentified genes.

Coding Strand Retrieval.—I identified template sequences for each gene by subtracting chicken subject start position from chicken subject end position for all BLASTn results using a custom python script. For those blast results with positive start minus end positions, this indicated that the sequenced DNA strand was the template, and the reverse compliment had to be retrieved to get the translated sequence. I reverse complimented 6555 sequences for which I only had the template strand out of 13,211 common loon sequences that met length criteria using a custom python script. Sequences on reverse versus forward strands were then sorted in Microsoft Excel to retain order. After coding sequences were obtained for all 13,211genes, I implemented an approach to check and adjust the reading frame where required. This was necessary because BLAST-identified [23] genes were fragments rather than entire coding sequences of mRNA transcripts. Thus, returned alignments of common loon and chicken gene fragments were created by BLAST [23] to optimize *E*-score based on nucleotide rather than codon alignment and could therefore start with nucleotides from incomplete codons. As such bases shift the reading frame, I used a custom python script to identify and correct open reading frames. First, I searched each nucleotide-aligned ortholog pair of common loon and chicken for multiple stop codons or stop codons in the middle of the sequence. Only modifying sequences that met these criteria, I iteratively added a maximum of two N bases to the beginning of each out of frame ortholog until stop codons disappeared. Frame adjustment removed 76 % of chimeric stop codons, but for the remaining 15% of sequences (3105) multiple stop codons remained. I attributed these to alignment gaps inserted by BLAST that do not make biological sense. Sequences for which the frame could not be corrected using these modifications were removed from the analysis.

Appendix 2. Description of Computer Resources

The following programs described in the content of the thesis above were run on a five-node compute cluster administrated by Dr. Jeff Horn:

ABySS- MPI

SAMTOOLS

BBMAP

BWA (SW)

PAML

The Williac cluster computer, located in the Cluster Laboratory of NMU's Mathematics and Computer Science Department in room 2106 of Jarnrich Hall, consists of ten rack-mounted computers. All nodes host Intel Core i7 quad-core processors with sixteen gigabytes of RAM and solid state drives. They are interconnected by a gigabit per second Ethernet switch and run Rockscluster software on top of the centOS Linux operating system. The ABySS runs took approximately eight to ten hours using five of the compute nodes.

Appendix 3. Locations of Genomic Assembly Data and Scripts

There are three main directories of files generated from this research:

(1): Genome Assembly.—The fasta formatted genome assembly produced with ABySS [16] using *k* = 30 is located in the directory COLO_Assembly along with supplemental ABySS [16]files.

(2): Scripts and Manipulation.—All Python language environment scripts used for parsing and analyzing the common loon genome assembly are located in the directory Loon-Scripts; descriptions of the script function are given as comments (#) in the script itself.

(3): Gene Data.—All files produced during genome annotation and analysis, including lists of identified genes, Gene Ontology associations, and values from selection analyses are located in the directory COLO_Supplemental_Files. Supplemental file 1 consists of a list of all 13,821 identified genes in the common loon, Supplemental file 2 consists of a list of Gene Ontology categories for all these identified genes, and Supplemental file 3 consists of results from evolutionary analyses.