

https://dspace.org/

The genome sequence of the Yellow-billed Duck (Anas undulata)

Ngxamani, Namhla

2024-01

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/38889

Ngxamani,N. (2024). The genome sequence of the Yellow-billed Duck (Anas undulata) [Masters dissertation, University of the Witwatersrand, Johannesburg]. https://hdl.handle.net/10539/38889

Downloaded from a DSpace repository. https://dspace.org/



The genome sequence of the Yellow-billed Duck (*Anas undulata*)

by

Namhla Ngxamani

(1653317)

Dissertation

Submitted in fulfilment of the requirements for the degree

Master of Science

in

Molecular and Cell Biology

in the Faculty of Science, University of the Witwatersrand, Johannesburg, South Africa

Supervisor: Dr Jean Mollett

Co-supervisor: Prof. Pieter De Maayer

January 2024

Declaration by the Student

I, Namhla Ngxamani (1653317), am a student registered for the degree of Master of Science in the academic year 2023.

I hereby declare the following:

- I am aware that plagiarism (the use of someone else's work without their permission and/or without acknowledging the original source) is wrong.
- I confirm that the work submitted for assessment for the above degree is my own unaided work except where explicitly indicated otherwise.
- I have not submitted this work before for any other degree or examination at this or any other University.
- I have followed the required conventions in referencing the thoughts and ideas of others.
- I understand that the University of the Witwatersrand may take disciplinary action against me if there is a belief that this is not my own unaided work or that I have failed to acknowledge the source of the ideas or words in my writing.

Signature: 

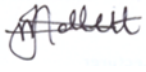
Signed on_12_ day of _January_2024

'Declaration by the Supervisor/s and Co-supervisor/s'

I declare that I have reviewed final version of the above-mentioned submitted work.

Supervisor

Name: Dr Jean Mollett

Signature: 

Signed on _12_ day of _January_2024

Co – supervisor

Name: Prof Pieter De Maayer

Signature: 

Signed on_12_day of _January_2024

To my late aunt and grandmother

Lindelwa Flatela

13 March 1967 – 15 July 2021

Nongezile Ngxamani

20 June 1940 – 30 September 2021

Acknowledgements

Thank you to Doctor Chevonne Reynolds, who kindly donated the blood samples that were used in this study. Thank you so much to my wonderful supervisors, Doctor Mollett and Professor De Maayer. Thank you for always demonstrating kindness and being available to assist me throughout my project.

Thank you, Doctor Botes, for having an open-door policy and serving as a sounding board for all my new ideas. Professor Kaur, my academic advisor, thank you for always making sure that I take care of my physical and mental health.

It was a difficult journey that would not have been possible without the support of my family. Thank you for always supporting me in any way, my mother (Zukiswa Ngxamani), big brother (Sihle Ngxamani), and new adopted big brother (Yanga Madikazi). I am truly blessed to call you family.

To my chosen family, thank you for all your continued love and support, Altron Godfrey, Keelan Krinsky, Nontuthuzelo Mlotshwa, Tanner Kerspuy, Xolani Mothopeng, and Zainab Nosarka. You aided me in more ways than I can count during this project. To my amazing talent manger Monty Meyer and very understanding supervisor Phuti Masenya, thank you so much for the unwavering support.

Finally, thank you to my sponsors, Gauteng City Region Academy (GCRA), Cipla Medpro, and the National Research Foundation (NRF), for allowing me to pursue my passion.

Table of Contents

List of figures i
List of tables iii
Abbreviations iv
Nomenclature vii
Chapter 1: The South African indigenous Yellow-billed Duck 1
Introduction 1
1.1 Yellow-billed Duck characteristics 1
1.2 South African Yellow-billed Duck ecological impact 3
1.3 Yellow-billed Duck status and conservation 5
1.4 Yellow-billed Duck threats 5
1.4.1 Hunting 6
1.4.2 Avian botulism 6
1.4.3 Mallard driven hybridisation 7
1.5 Project rationale and aim 14
Chapter 2: Sequencing, *de novo* assembly and annotation of the first Yellow-billed Duck (*A. undulata*) genome 16
2.1 Introduction 17
2.2 Method 18
2.2.1 Sampling 18
2.2.2 Whole genomic DNA extraction and integrity assessment 19
2.2.3 PacBio continuous long read (CLR) and high-fidelity (HiFi) read genome sequencing 20
2.2.4 *De novo* long read genome assembly with Canu 24
2.2.5 Yellow-billed Duck genome annotation 28
2.3 Results and Discussion 33
2.3.1 Yellow-billed Duck genome raw data quality assessment 33
2.3.2 *De novo* assembly 38
2.3.3 Analysis of adapter contamination 39

2.3.4 Genome completeness analysis 40
2.3.5 Repeat element identification and annotation 40
2.3.6 RNA prediction and annotation 41
2.3.7 *Ab initio* gene prediction output 43
2.3.8 Functional annotation 44
2.4 Conclusion 48
Chapter 3: Comparative genomics of the Yellow-billed Duck 49
3.1 Introduction 50
3.1.1 *Anas* species to avian genomic analysis 51
3.2 Methods 52
3.2.1 Assembly comparative analysis 52
3.2.2 Annotation comparative analysis 53
3.2.3 Orthology inference 53
3.2.4 *Anas* core genome phylogeny 53
3.2.5 Pangenome analysis 53
3.2.6 Paralogue analysis 54
3.3 Results and discussion 55
3.3.1 Assembly comparative analysis 55
3.3.2 *Anas* core genome phylogeny 58
3.3.3 Annotation comparative analysis 61
3.3.4 Repeat element comparative analysis 62
3.3.5 Pangenome analysis 64
3.3.6 Paralogue analysis 65
3.4 Conclusion 73
Chapter 4: Summary and concluding remarks 74
4.1 Project summary 75
References 78
Appendix 100

List of figures

Figure 1.1: Photograph showing the phenotype of the monomorphic Yellow-billed Duck.

Figure 1.2: Current Yellow-billed Duck and Mallard distribution and sampling location.

Figure 1.3: Photographs showing dimorphic Mallard Ducks and the foot of a proposed, Yellow-billed Duck × Mallard hybrid.

Figure 2.1: PacBio long-read sequencing framework.

Figure 2.2: PacBio Sequel IIe continuous long read (CLR) and high-fidelity (HiFi) sequencing mechanism.

Figure 2.3: *De novo* long-read assembly framework. Grey blocks indicate assembly datasets, and the red line indicates the path to the final *de novo* assembly.

Figure 2.4: MOSGA pipeline sequential software database utilisation and subsequent file outputs for the primary genome.

Figure 2.5: CLR and HiFi read length distribution.

Figure 2.6: Quality score (Q) read distribution.

Figure 2.7: GC content distribution across total CLR and HiFi reads.

Figure 2.8: BUSCO genome completeness assessment.

Figure 2.9: Distribution of transfer RNA (tRNA) genomic hits from 260 tRNA isotypes predicted on the YBD genome.

Figure 2.10: Classification of Yellow-billed Duck gene products into the 26 COG functional categories according to EggNOG-mapper alignment against the eggNOG 5.0 database.

Figure 2.11: Classification of Yellow-billed Duck gene products into the 26 COG functional categories according to DIAMOND alignment against the SwissProt database.

Figure 2.12: The Yellow-billed Duck's enzymatic profile from 4,503 KEGG hits.

Figure 3.1: Genome assembly completeness assessment using BUSCO v.4.1.4.

Figure 3.2: Core genome phylogeny of eight ducks in the genus *Anas*.

Figure 3.3: Midpoint rooted maximum likelihood phylogeny of the genus *Anas*.

Figure 3.4: Repeat element genomic composition across nine respective genomes.

Figure 3.5: The genus *Anas* pangenome, with the *G. gallus* as the outgroup.

Figure 3.6: Paralogous genes of wild *Anas* genomes.

Figure 3.7: Paralogous comparative analysis of the YBD and domestic *Anas* genomes.

Figure 3.8: Paralogous comparative analysis of wild and domestic *Anas* genomes.

Figure 3.9: Species-specific RCFV heatmap.

List of tables

Table 2.1: *De novo* assembly metrics assembled using Canu v.2.2.

Table 2.2: VecScreen data output summary.

Table 2.3: Yellow-billed Duck repeat element genomic composition.

Table 2.4: AUGUSTUS v.3.4.0 output summary.

Table 3.1: Assembly statistics of the Yellow-billed Duck and comparative genomes.

Table 3.2: The gene, functional annotation, tRNA, and rRNA genome summaries for all comparative Aves.

Table A1: Experimental data summary of the Yellow-billed Duck and comparative genomes.

Abbreviations

AEWA	Agreement on the Conservation of African Eurasian Migratory Waterbirds
BOG	Best Overlap Graph
BoNT	Botulinum Neurotoxin
BUSCOs	Benchmarking Universal Single Copy Orthologs
CCS	Circular Consensus Sequencing
CLR	Continuous Long Read
COG	Clusters of Orthologous Group
DAG	Directed Acyclic Graph
DDBJ	DNA Data Bank of Japan
EDTA	Ethylenediamine tetraacetic acid
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
GFA	Graphical Fragment Assembly
GHMM	Generalised Hidden Markov Model
GIAB	Genome in a bottle
GO	Gene Ontology
GUI	Graphical User Interface
HiFi	High Fidelity
HMM	Hidden Markov Model
IBA	Important Bird and Biodiversity Area
IMM	Interpolated Markov Model

KEGG	Kyoto Encyclopedia of Genes and Genomes
LINE	Long Interspersed Nuclear Element
LTR	Long Tandem Repeat
MHAP	MinHash Alignment Process
MOSGA	Modular Open-Source Genome Annotator
MSA	Multiple sequence alignment
NCBI	National Center for Biotechnology Information
NGS	Next-Generation sequencing
NMJ	Neuromuscular Junction
OBT	Overlap-Based Trimming
OLC	Overlap Layout Consensus
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PCI	Phenol:Chloroform:Isoamylalcohol
PSMC	Pairwise Sequentially Markovian Coalescent (PSMC) technique
QLB	Queens lysis buffer
RQ	Read quality score
SAWA	Southern African Wingshooters Association
SCO	Single Copy Ortholog
semi-CRF	semi-Markov Conditional Random Field
SGS	Second-Generation Sequencing
SINE	Short Interspersed Nuclear Element

SMRT	Single-Molecule Real-Time
SNARE	Soluble N-Ethylmaleimide Sensitive Factor Attachment Protein Receptors
SNP	Single Nuclear Polymorphism
SPOA	Single Instruction Multiple Data (SIMD) and Partial-Order Alignment (POA)
STR	Short Tandem Repeat
TBE	Tris aminomethane (Tris), Boric acid and EDTA
TGS	Third-generation sequencing
tf-idf	term frequency, inverse document frequency weighing
WWAM	Windowed Weight Array Model
YBD	Yellow-billed Duck
ZMW	Zero-Mode Waveguide

Nomenclature

Nomenclature	Definition
Circular consensus sequencing (CCS)	PacBio's method of aligning multiple overlapping reads from a circular DNA molecule and determining the most likely base at each position (Rhoads & Au, 2015, Hon *et al.*, 2020; Logsdon *et al.*, 2020).
Consensus sequence	Bases that appear with highest frequency at each position when a series of sequences believed to have common function is compared (Garrett & Grisham, 2016).
Containments	Overlaps in which all bases of one read align with the bases in another, or dovetail, if only the read ends are involved in the alignment (Koren *et al.*, 2017)
Contig(s)	Contiguous stretch(es) of DNA sequence without gaps that have been assembled solely based on direct sequencing information (Logsdon *et al.*, 2020).
Dimorphism	The condition of occurring in or representing two distinct forms within a given species and traits that differ consistently between males and females (Kliman, 2016).
Ecosystem services	Organism-based activities indirectly or directly beneficial to humans for their physical and socioeconomic well-being (Sekercioglu, 2006).
Ectozoochory	Seed dispersal via the fur or skin of animals (Sekercioglu, 2006; Suri *et al.*, 2017).
Endozoochory	Seed dispersal via ingestion by vertebrate animals (Sekercioglu, 2006; 2016; Suri *et al.*, 2017).
Functional group	A primarily diet-based grouping that parallels the main ecological service of the species (Sekercioglu, 2006).
Genetic differentiation	The accumulation of differences in allelic frequencies between completely or partially isolated populations due to evolutionary forces (McGirr & Martin, 2020).
Genetic incompatibility	Interactions between alleles that result in lowered fitness of individuals carrying their combination (Kulmuni & Westram, 2017).

Heterospecific gene flow	The movement of genetic material from one species population into another different species population (Mallet, 2005; Kliman, 2016).
Holarctic	Geographical region comprising of the Nearctic (subtropical, tropical, temperate, and arctic North America) and Palearctic (Europe, northern Africa, and Asia north of the Oriental region) regions (Katinas *et al.*, 1999).
Hub species	Nodes (species) in a hybrid network (interactions between several hybridising species) have the most interactions and so hybridise with a large number of different other species (Ottenburghs, 2019).
Hybridisation	The interbreeding of two genetically variant species, resulting in the production of hybrids (Harrison & Larson, 2014; de Souza *et al.*, 2019; Stephens *et al.*, 2019).
Intergeneric hybridisation	Hybridisation between species in two different genera in the same family (Grant & Grant, 1992; Ottenburghs *et al*., 2015; Ottenburghs, 2019; Lawson *et al.*, 2021).
Interspecific hybridisation	Hybridisation of two species from the same genus (Grant & Grant, 1992; Ottenburghs *et al*., 2015; Lawson *et al.*, 2021).
Introgressive hybridisation	The spread of genes of one species into the gene complex of another because of hybridisation between numerically dissimilar populations in which extensive backcrossing prevents formation of a single stable population (Anderson, 1948; Harrison & Larson, 2014).
Isoform	Any two or more functionally similar proteins with a similar but not identical amino acid sequence and are either encoded by different genes or by RNA transcripts from the same gene which have had different exons removed (Garrett & Grisham, 2016).
Light pulse	White light with a high intensity that is emitted in a series of brief flashes (Eid *et al.*, 2009; Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020).
Long-read sequencing	PacBio and Oxford Nanopore Technologies (ONT) use a sequencing method in which target DNA or RNA molecules are sequenced in real time, often without the need for amplification, producing long (>10 kb) reads (Logsdon *et al.*, 2020).

Mobile link	Organisms that actively move in the landscape and connect habitats in space and time (Sekercioglu, 2006).
N50	The length of the shortest contig for which longer and equal length contigs cover at least 50% of the assembly (Alhakami *et al.*, 2017).
Next-generation sequencing	A low-cost, rapid, and large-scale sequencing method that produces high-quality sequencing reads from genetic material fragments (Goffeau *et al*., 1996; Shendure *et al.*, 2017).
Orthogroup	A set of genes descended from a single gene of the last common ancestor of in all the species being evaluated (Emms & Kelly, 2015).
Ortholog	Genes in different species that evolved from a common ancestral gene by speciation (Kliman, 2016).
Parsimony informative site	A site that contains no less than two or more-character states in two or more taxa (Kumar *et al.,* 2016).
Phylogenetic inference	The practice of reconstructing the evolutionary history of related species by grouping them in successively more inclusive sets based on shared ancestry (Minh *et al*., 2020).
Poisson read distribution	A discrete probability distribution for the number of events that occur in each observational time (Myers, 2014).
Propagule	Any cellular structure produced by an organism that is capable of dispersing and surviving in the environment before developing into a new individual (Sekercioglu, 2006; Reynolds, 2016; Suri *et al.*, 2017).
Single-molecule, real-time (SMRT) sequencing	PacBio's DNA sequencing method in which a single DNA molecule is sequenced and derived in real time, with no pause after the bases are detected (Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020).
Singleton site	An alignment site that contains at least two types of nucleotides or amino acids with, at most, one occurring multiple times (Kumar *et al.,* 2016).
SMRT 8M Cells or Zero-Mode Waveguide (ZMW) units	Nanophotonic chambers utilised in PacBio sequencing that confines light to a small observational volume (Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020).

SMRTbell molecule	PacBio DNA library molecule, consisting of a DNA insert and two pair-end ligated adapter hairpins (Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020).
Unitig	A unique subset of overlapping fragments that can be assembled. This means that a unitig is either a correctly assembled portion of a contig or it is an over-compressed assembly of several high-fidelity copies of a repeat (Miller *et al.*, 2008; Al-Okaily, 2016).

Chapter 1: The South African indigenous Yellow-billed Duck

Introduction

Within the amniotes clade, which includes birds, reptiles, and mammals, Aves have the highest species diversity (Ksepka, 2021) with 10,933 different species currently described (Gill *et al.*, 2021). South Africa is one of the avian biodiversity hotspots, housing ~7% of the global avian population (762 species), eighteen of which are endemic to the country, with the grassland and fynbos biomes having the highest levels of endemism, each supporting approximately 40.9% of the total avifauna (Wright & Lee, 2017; BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). The grassland biome serves as a preferred habitat for the 158 waterfowl inhabitants, with the biome acting as a water catchment area with less than 20% tree cover (Suri *et al.*, 2017; BirdLife International, 2022). Grassland filtered water bodies coupled with grassland comparable insect diversity and elevated phytodiversity of the fynbos create favourable habitat conditions for the inhabitation of endemic omnivorous waterfowl species (Procheş & Cowling, 2006; Suri *et al.*, 2017; Wright & Lee, 2017). This includes *Anas undulata* (Yellow-billed Duck), the focal species of this study, which had the highest number of southern African recoveries in one of the earliest studies on southern African waterfowl migration (Oatley & Prys-Jones, 1986).

1.1 Yellow-billed Duck characteristics

Yellow-billed Ducks (YBDs) are predominantly nocturnal omnivorous from the genus *Anas* within the order Anseriformes and class Aves (Livezey, 1991). *Anas undulata* has two subspecies: *A. undulata undulata* (southern YBD), native to eastern and southern Africa and *A. undulata ruppelli* (northern YBD; Blyth, 1855), concentrated in Ethiopia and Kenya (Woodman, 1944; Oatley & Prys-Jones, 1986). It should be noted that the focal species is the southern YBD, which will be referred to as the YBD in this work. Ducks of both subspecies are sexually monomorphic with a dark head, a distinctly recognisable yellow bill, grey-black plumage with a unique visible teal speculum, and yellow eyes (Figure 1.1; Kear, 2005; de Souza *et al.*, 2019). Males and females have similar plumage, but males are larger than females and both have grey-black legs and feet (Figure 1.1; Kear, 2005; de Souza *et al.*, 2019; Southern African Bird Atlas Project 2, 2022). YBDs weigh between 700 and 1,150 g, with height ranging between 52 and 58 cm (Kear, 2005; de Souza *et al.*, 2019; Southern African Bird Atlas Project 2, 2022). They have a lifespan of between 20 and 30 years (BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022).



Figure 1.1: Photograph showing the phenotype of the monomorphic Yellow-billed Duck (Photograph by S. Carreavesra).

YBDs are dispersive and nomadic birds with a low migration rate extensively distributed throughout South Africa, with the largest population density seen in the above-mentioned South African biomes illustrated in Figure 1.2 (Oatley & Prys-Jones, 1986; Maclean, 1997; Brown *et al.*, 2019; de Souza *et al.*, 2019; BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). The YBD habitat is a body of water with a pH greater than 10, such as pools created by laminar flowing rivers adjacent to floored grasslands, streams, lakes, marshes, sewage, sewage works, and mining-linked manufactured reservoirs (BirdLife International, 2022). The species is monogamous, with a year-round breeding period that peaks from June to November in the southern regions of Western Cape and from December to April in northern parts of the country, facilitated by seasonal wetland migration (Maclean, 1993; Little *et al.*, 1995; Southern African Bird Atlas Project 2, 2022). The courtship behaviour of the species involves emitting a grunt whistle and performing head-up-tail-up and down-up courtship displays (Young, 1999).

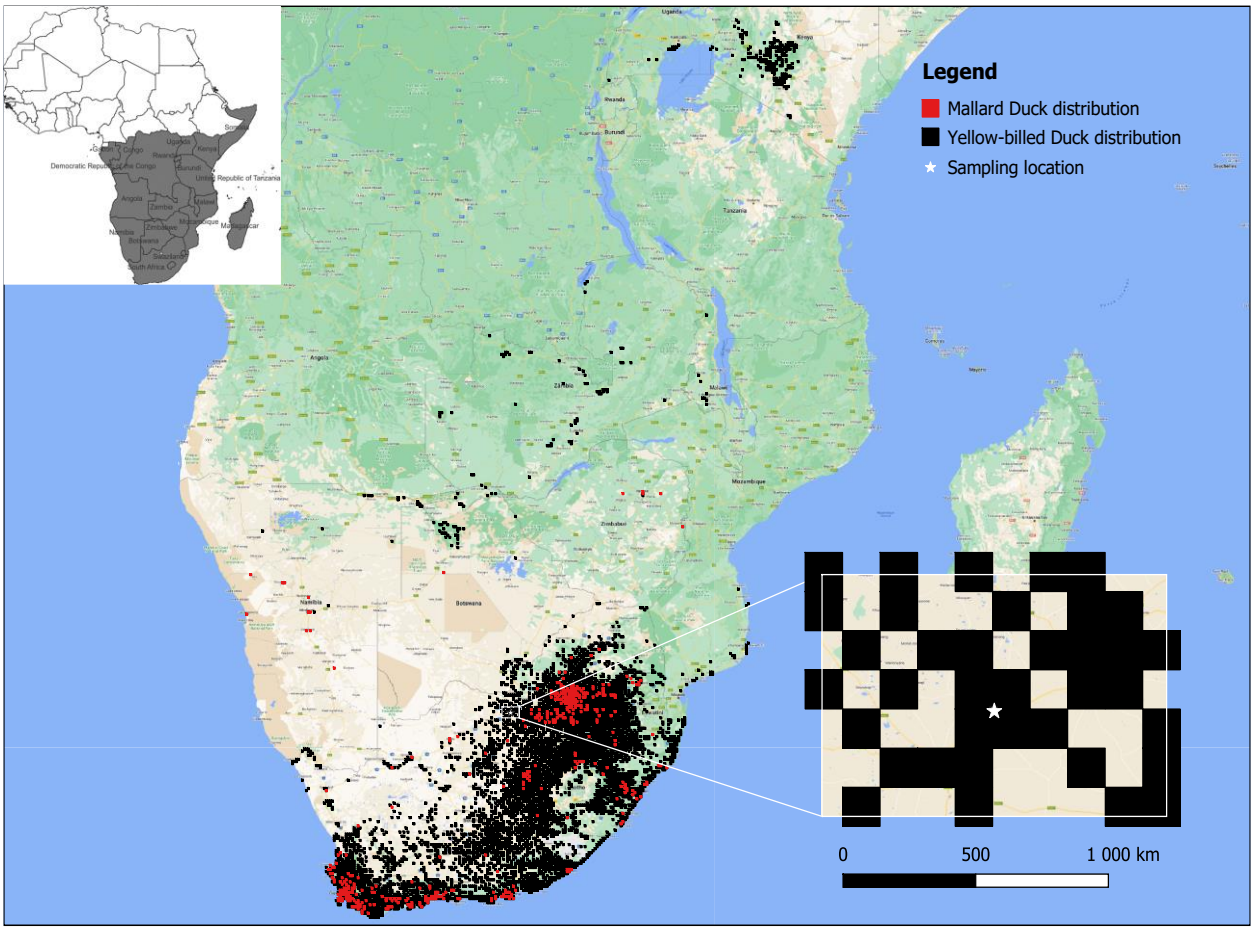


Figure 1.2: Current Yellow-billed Duck and Mallard distribution and the sampling location. Within the zoomed-in region, a white star indicates the sampling location. Distribution data based on phenotypic expression. Data from the Southern African Bird Atlas Project 2 (2022).

1.2 South African Yellow-billed Duck ecological impact

In addition to being part of the indigenous South African avifauna, the YBD population is of great ecological importance due to its ability to provide ecosystem services and connect isolated habitats by acting as a mobile link (Sekercioglu, 2006; Reynolds, 2016). Ecosystem services are organism-based activities that are directly or indirectly beneficial to humans, with avitourism providing a direct anthropogenic benefit to the South African economy, contributing ~R1.725 billion in 2010 and expected to grow each year (Sekercioglu, 2006; Department of Trade and Industry, 2010; Suri *et al.*, 2017). These services are divided into different ecological functional groups such as nutrient deposition, aquatic propagule dispersal, avian pollination, and insect regulation facilitated by the omnivorous diet of the species (Sekercioglu, 2006; Reynolds, 2016; Suri *et al.*, 2017).

Furthermore, waterfowl activity, such as that of the YBD, delivers a considerable amount of nutrients to isolated habitats, accounting for 40% of total nitrogen and 75% of phosphorus deposited into wetlands (Sekercioglu, 2006).

The YBD serves as a propagule vector as its diet of aquatic insects and their larvae, as well as plant material from fruits and seeds, promotes not only phyto-propagule germination within avian guts but also phyto-propagule and macroinvertebrate dispersal via endo- and ectozoochory (Sekercioglu, 2006; Damschen *et al.,* 2008; Holyoak *et al.,* 2008; Reynolds, 2016; Suri *et al.,* 2017). Studies of multiple, YBD endo- and ectozoochoric samples for dormant aquatic invertebrate eggs in Barberspan discovered that fifteen of 60 endozoochoric and five of 95 ectozoochoric samples contained *Daphnia ephippia*, a species of water flea, while ten of the 95 endozoochoric samples contained endemic *Lophopodella capensis* (a marine gastropod crustacean) statoblasts (Reynolds & Cumming, 2015; Reynolds, 2016). *Daphnia* spp. are of great ecological significance in wetlands as species activities maintain food chain energy transfer, provide algal top-down biocontrol, and serve as a bioindicator (Tomasiks & Warren, 1996; Govaert *et al.,* 2021).

Elevated flora biodiversity within the isolated habitat of the fynbos indicates the presence of mobile organisms interconnecting remote ecosystems via passive dispersal (van Leeuwen *et al.,* 2012). Localised movement of Anseriformes across the South African landscape promotes ecto- and endozoochoric (requires gastrointestinal tract resistance) phyto-propagules and macroinvertebrate dispersals, resulting in species introduction and the maintenance of species biodiversity in isolated habitats (van Leeuwen *et al.,* 2012; Reynolds, 2016; Suri *et al.,* 2017; BirdLife International, 2022). Mobile links, defined as a migratory organism whose actions connect isolated inhabitants in space and time, link various habitats through nutrient provision and propagule dispersal (Reynolds, 2016; Suri *et al.,* 2017). It should be noted however, that selection and propagule dispersion is unregulated, which may contribute to the introduction of regional alien and invasive species. One investigation conclusively demonstrated this, identifying 66 intact ectozoochoric propagules and twenty-two intact endozoochoric propagules from 155 YBDs belonging to eleven plant species from five plant families, six of which were assumed alien to Barberspan (Reynolds, 2016).

1.3 Yellow-billed Duck status and conservation

Despite their migratory status, the native *A. u. undulata* is protected by the Agreement on the Conservation of African-Eurasian Migratory Waterbirds (AEWA), an international agreement put in place for migratory waterfowl conservation and their habitats (Oatley & Prys-Jones, 1986; Department of Environmental Affairs and Tourism of the Republic of South Africa, 2003; Brown *et al.*, 2019; de Souza *et al.*, 2019; BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). The AEWA highlighted the internationally significant population size of the South African YBD. The treaty also recognises the significance of South African Important Bird and Biodiversity Areas (IBAs) in providing suitable habitats and enhanced protection to a vast number of migratory birds like Baillon's Crake (*Porzana pusilla*), Crowned Crane (*Balearica regulorum*), Wattled Crane (*Bugeranus carunculatus*) and the YBD (Department of Environmental Affairs and Tourism of the Republic of South Africa, 2003).

Although the YBD is distributed across eastern and southern African countries, all five IBAs designated by BirdLife International for the species are in South Africa, with one of them serving as the sample location for this study (Barberspan) (Donald *et al.*, 2019; BirdLife International, 2022). IBAs are areas that house significant populations of threatened and endemic avian populations, identified based on a set of internationally accepted standards to aid in localised avian conservation for the preservation of global avian biodiversity (Donald *et al.*, 2019; Wright & Lee, 2017; BirdLife International, 2022). BirdLife International selected the IBAs based on the B2a regional IBA selection criterion, which states that a species has a positive conservation status and is concentrated within that region, and thus site-conservation within that region can aid in overall species conservation (BirdLife International, 2022). Despite the high South African localisation, the YBD population is still threatened by several factors.

1.4 Yellow-billed Duck threats

Conservation of the South African YBD is an important concern for the future of the species. Overhunting, botulism, and Mallard-driven hybridisation have been identified as potential threats to the species (BirdLife International, 2022).

1.4.1 Hunting

Avian game hunting (wingshooting) was established in ancient Egypt as a means of nourishment and survival (Geldenhuys *et al.,* 2013). Game hunting became a leisure activity around the world after the European invention of shotguns, causing a conservation issue due to overhunting (Bennett & Whitten, 2003; Geldenhuys *et al.,* 2013; BirdLife International, 2022). Overhunting is linked to the extinction of 76 bird species (47.23%) out of a total of 160 extinct birds since the early 1500s (Bennett & Whitten, 2003; Geldenhuys *et al.,* 2013, Gill *et al.,* 2021; BirdLife International, 2022). Waterfowl have accounted for 5% (eight) of all extinct Aves since the 1500s (Gill *et al.,* 2021; BirdLife International, 2022). Wingshooting is not just a threat to avifauna, but it can also undermine wetland integrity by removing birds that serve as ecosystem engineers and as previously noted, provide ecosystem services (Bennett & Whitten, 2003). AEWA, Department of Environmental Affairs, (2015) and BirdLife International (2022) have implemented hunting restrictions to protect the species. Despite AEWA protection, the Southern African Wingshooters Association (SAWA) still lists the YBD as gamefowl, with an average of 7.3 birds permitted to be hunted without a licence per person per day. AEWA protection based on species categorisation entails the enforcement of sustainable hunting practices (Department of Environmental Affairs of the Republic of South Africa, 2015).

1.4.2 Avian botulism

Avian botulism, identified in 1910 and confirmed in 1930, is the most potent waterfowl and shorebird disease, killing > 50 000 individuals per outbreak (Johnson & Bradshaw, 2001; Rocke & Bollinger, 2007; Hill & Smith, 2012; Govender *et al.,* 2019). It is caused by the ingestion of a proteinaceous botulinum neurotoxin (BoNT) secreted by *Clostridium botulinum* strains classified according to toxin subtypes ranging from A to G (Johnson & Bradshaw, 2001; Rocke & Bollinger, 2007; Hill & Smith, 2012; Govender *et al.,* 2019). Most waterfowl botulism outbreaks are caused by type C, however outbreaks caused by type E, or a protein mosaic of types C and D have also been documented (Rocke & Bollinger, 2007; Govender *et al.,* 2019).

C. botulinum inhabits aquatic sediments, like wetlands and shorelines where avian botulism is most prevalent (Johnson & Bradshaw, 2001; Rocke & Bollinger, 2007; Govender *et al.,* 2019). These regions are anoxic with warm temperatures, making them favourable for the formation of

C. botulinum populations and the generation of BoNT as a secondary metabolite (Govender *et al.*, 2019). BoNTs function as sequence specific endopeptidases, cleaving SNAREs (soluble N-ethylmaleimide sensitive factor (NSF) attachment protein receptors), preventing the release of acetylcholine at neuromuscular junctions (NMJs) (Gu & Jin, 2012). This causes flaccid paralysis, which manifests as fatigue, muscle weakness and a limber neck, with death usually resulting from drowning or respiratory failure (Gu & Jin, 2012; Govender *et al.*, 2019).

Avian botulism is the leading cause of waterfowl death in North America; however, the disease is less prevalent in South Africa, with the 2017-2018 H5N8 viral infection cited as the most recent waterfowl outbreak in the country (Rocke & Bollinger, 2007; Govender *et al.*, 2019; Department of Forestry, Fisheries, and the Environment of the Republic of South Africa, 2022). Govender *et al.* (2019) identified 9% of the 1,115 botulism-affected waterfowl carcasses investigated between 2015 and 2017 as belonging to the YBD, with the Red-knobbed Coot (*Fulica cristata*) having the highest mortality rate at 60%. Continuous monitoring of the disease in shorebirds and waterfowl in warm anoxic aquatic regions is required to ensure low outbreak levels (Govender *et al.*, 2019). Although no species extinctions have been connected to botulism, *Anas laysanensis* (Laysan Duck) is categorised as critically endangered, owing to an initial 2008 type C botulism outbreak that killed 160 individuals and recurring outbreaks each year thereafter (BirdLife International, 2022).

1.4.3 Mallard driven hybridisation

1.4.3.1 Sympatric introgressive hybridisation

Hybridisation is defined as the interbreeding of two genetically variant species, resulting in the production of hybrids (Harrison & Larson, 2014; de Souza *et al.*, 2019; Stephens *et al.*, 2019). According to Hubbs' principle (1955), hybridisation is most prevalent in regions where anthropogenic activity has influenced physical barrier removal, resulting in the coexistence of related species where the population density of one species is lower (Grant & Grant, 1997; Randler, 2002). Hybridisation is one of the most detrimental effects caused by invasive species (Rhymer & Simberloff, 1996; Harrison & Larson, 2014; de Souza *et al.*, 2019; Ottenburghs, 2019; Stephens *et al.*, 2019). Fertile hybrid progeny production and continuous backcrossing may allow for continuous gene flow post hybridisation, resulting in the permanent integration of genes from one gene pool into another, a process known as introgression (Rhymer & Simberloff, 1996; Stephens

et al., 2019). Introgressive hybridisation is defined as the integration of genes from one organism into the genome of another, resulting in stable genomic regions of the in- or crossbred species that may be identified as different genome pools (Anderson, 1948; Harrison & Larson, 2014).

The occurrence in which related species rejoin and produce genetically different offspring populations from the parents (secondary contact), occurs in approximately 25% of plants and 10% of animals, and is essential for shaping the evolutionary paths of various plants and animals (Rhymer & Simberloff, 1996; Mallet, 2005; Genovart, 2009; McCracken & Wilson, 2011). Amongst vertebrates, this occurs most frequently in avian species, in 2015 hybridisation had been documented in 16% (1,714) known 10,446 bird species (Grant & Grant, 1992; Genovart, 2009; McCracken & Wilson, 2011; Ottenburghs *et al*., 2015). Avian species routinely hybridise with just one other species, but hybridisation of multiple species within the 44 avian orders does occur (Ottenburghs, 2019; Lawson *et al.*, 2021; Gill *et al.*, 2021). The highest level of avian hybridisation observed through morphological analysis is within the order Anseriformes, with approximately 41.6% of 161 anseriform species undergoing interspecific hybridisation and 30.7% undergoing intergeneric hybridisation (Grant & Grant, 1992; Ottenburghs *et al*., 2015; Lawson *et al.*, 2021).

1.4.3.2 Most prolific hub species: The Mallard Duck

The Mallard Duck (*Anas platyrhynchos*) comprises three currently recognised subspecies, namely *A. platyrhynchos domesticus*, *A. platyrhynchos conboschas* and *A. platyrhynchos platyrhynchos* (BirdLife International, 2022; Gill *et al.*, 2021; Southern African Bird Atlas Project 2, 2022). The common domestic duck is the *A. p. domesticus*, with the Pekin Duck domesticated primarily for meat production and the Laying Duck domesticated primarily for egg production (Huang *et al.*, 2013; Zhang *et al.*, 2018). Domestic ducks are important agricultural and economic commodities in Southeast Asia, particularly southern China, which has been a major focal point of duck domestication for at least 2,200 years (Serjeantson, 2014; Zhang *et al.*, 2018; Wang *et al.*, 2020). The Mallard has a global distribution (Gill *et al.*, 2021). Mallards originate from the Northern Hemisphere, with a native holarctic distribution and an estimated global population size of over 19 million individuals (Madge & Burn, 1988; Rhymer & Simberloff, 1996; Katinas *et al.*, 1999; Brown *et al.*, 2019; de Souza *et al.*, 2019; BirdLife International, 2022).

Both wild sexes of Mallard have orange legs and feet, a mass ranging from 850 to 1,400 g, and a height ranging from 50 to 65 cm (Madge & Burn, 1988; de Souza *et al.*, 2019; Southern African Bird Atlas Project 2, 2022). Specific sex-based phenotypes are depicted in Figure 1.3 (Madge & Burn, 1988; de Souza *et al.*, 2019; Southern African Bird Atlas Project 2, 2022). Most *A. p. domesticus* sexes are monomorphic, with a distinctly white plumage and yellow bill phenotype being more prevalent (Serjeantson, 2014; Zhang *et al.*, 2018). Because these ducks are desired for meat and egg production, the skeleton of the common domestic duck shows that it is now too heavy to fly, with the Pekin Duck weighing 5,500 g (Serjeantson, 2014; Zhang *et al.*, 2018). Pekin Ducks, unlike wild Mallards and YBDs, prefer to perform the down-up courtship display rather than the grunt-whistle or head up-tail up courtship displays (Miller, 1977; Young, 1999).

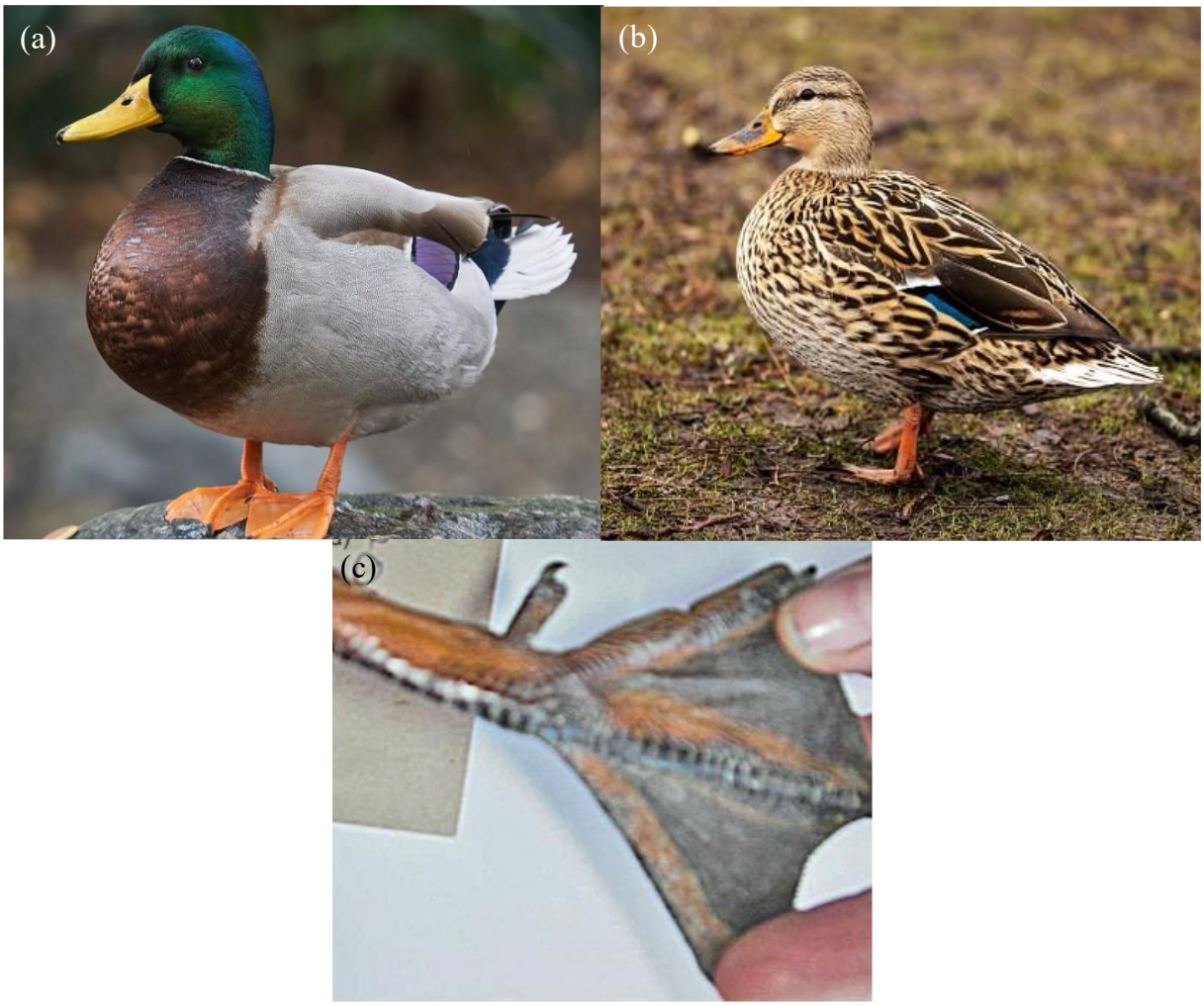


Figure 1.3: Photographs showing dimorphic Mallard Ducks and the foot of a proposed, Yellow-billed Duck × Mallard hybrid. (a) and (b) illustrate the sexual dimorphism of the Mallard species

with (a) male (Photograph by G. Heaton) and (b) female birds (Photograph by D. Johnston). (c) leg and foot of an inferred, Yellow-billed Duck × Mallard hybrid (Photograph by C. Reynolds, adapted from de Souza *et al.* (2019)).

1.4.3.3 Mallard driven hybridisation

Some of the best characterised hybridisation events are those occurring in dabbling ducks facilitated by the accidental or deliberate introduction of the Mallard Duck (Grant & Grant, 1992; Rhymer *et al.*, 1994; Rhymer & Simberloff, 1996; Rhymer, 2006; Muller, 2008; Genovart, 2009; Ottenburghs *et al*., 2015; de Souza *et al.*, 2019; Ottenburghs, 2019; Stephens *et al.*, 2019; Lawson *et al.*, 2021). The Mallard being the most prolific hub species, has formed hybrids with at least 39 to 60 distinct species within the order Anseriformes (McCarthy, 2006; de Souza *et al.*, 2019; Ottenburghs, 2019). Mallards can undergo interspecific hybridisation as well as intergeneric hybridisation with distant species such as *Netta rufina* (Red-crested Pochard) and *Anser anser* (Greylag Goose) (Ottenburghs, 2019).

The introduction of the Mallard to countries such as Australia, New Zealand, South Africa and the United States has resulted in native species population declines in Mallard related species, which have been attributed to introgressive hybridisation (Rhymer &Simberloff, 1996; Rhymer, 2006; Muller, 2008; de Souza *et al.*, 2019; Stephens *et al.*, 2019). For example, Mallard hybridisation is linked to the loss of genetic variation in *Anas rubripes* (American Black Duck) (Rhymer & Simberloff, 1996; Rhymer, 2006; Brown *et al.*, 2019; de Souza *et al.*, 2019). Agriculturally influenced habitat modification resulted in sympatric introgressive Mallard hybridisation with *Anas superciliosa superciliosa* (New Zealand Grey Duck) and caused a population decline of the latter duck (Rhymer & Simberloff, 1996; Rhymer, 2006; Muller, 2008; Brown *et al.*, 2019; de Souza *et al.*, 2019; Lawson *et al.*, 2021). Morphological data indicate that Mallard Ducks can hybridise with various southern African avian species, most noteworthy being *Anas smithii* (Cape Shoveler), *A. capensis* (Cape Teal), *A. sparsa* (African Black Duck) and the YBD (de Souza *et al.*, 2019; Stephens *et al.*, 2019).

The Mallard was introduced to the Western Cape for hunting purposes in the 1940s, and the population rapidly became prevalent (Brown *et al.*, 2019; de Souza *et al.*, 2019; Stephens *et al.*, 2019). Mallards are currently primarily found in the northeastern Gauteng Province and the

southern parts of the Western Cape, with recorded sightings throughout the country as depicted in Figure 1.2 (Southern African Bird Atlas Project 2, 2022).

1.4.3.4. Mallard and YBD hybridisation

The proximity and greater population size of the Mallard relative to the YBD population raises concerns regarding YBD genetic integrity. This concern is based on Hubbs' principle, which states that "great scarcity of one species coupled with the abundance of another often leads to hybridisation: the individuals of the sparse species seem to have difficulty in finding their proper mates" (Hubbs, 1955; Oatley & Prys-Jones, 1986; McCracken & Wilson, 2011; Brown *et al.*, 2019; de Souza *et al.*, 2019; BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). Possible Mallard × YBD hybrids, are thought to be fertile and have been identified in five of the nine provinces (Eastern Cape, Free State, Gauteng, KwaZulu-Natal, and Western Cape) (de Souza *et al.*, 2019; Stephens *et al.*, 2019; Southern African Bird Atlas Project 2, 2022). The hybrids have a YBD phenotype with Mallard-like orange legs and feet (Figure 1.3c; de Souza *et al.*, 2019; Stephens *et al.*, 2019; Southern African Bird Atlas Project 2, 2022). Mallards have since been classified as group 2 invasive species under the National Environmental Management Biodiversity Act (10/2004): Alien and Invasive Species Regulations (Government of the Republic of South Africa, 2014), which means the keeping of Mallard Ducks requires that the ducks are kept in a controlled area and the owner requires a permit.

1.4.3.5 Physiological Mallard characteristics supporting hybridisation

Mallard driven hybridisation is reported to result in significant adverse genetic impacts, interspecific population reductions, and localised extinction events (Rhymer & Simberloff, 1996; McCarthy, 2006; de Souza *et al.*, 2019). Typically, Mallards are larger than other *Anas* species, are more promiscuous (with a longer breeding window) and aggressive (Williams & Basse, 2006). These properties drive forced copulation of other ducks, thus resulting in hybridisation through the forced copulation mechanism (Williams & Basse, 2006; de Souza *et al.*, 2019). These behavioural attributes can be linked to increased testosterone production in domesticated Mallard drakes as seen by Paulke and Haase (1978). Wingfield *et al.* (1987), highlighted that the expression of domineering behaviour is usually linked to increased aggression and elevated testosterone concentrations. Furthermore, the elaborate feathers of Mallard drakes have been proposed to

stimulate other hen species and promote hybridisation (de Souza *et al.,* 2019). Additionally, the great similarity of the species' mating displays described by Young (1999), can potentially aid in Yellow-billed Duck × Mallard hybridisation.

1.4.3.6 Ramifications of hybridisation

Based on genetic and phenotypic studies Mallards have subsequently been ruled invasive to New Zealand, Australia, and South Africa (de Souza *et al.*, 2019; Stephens *et al.*, 2019). Anthropogenic activity has accelerated the introduction of biological invasive species (Rhymer & Simberloff, 1996; Mallet, 2005), resulting in biodiversity loss. Although natural hybridisation has been documented, it is attributed to environmental alterations (Grant & Grant, 1992; Grant & Grant, 2019).

If Hubbs' principle requirements are met, the erosion of physical barriers that result in sympatry and subsequent introgressive hybridisation can also result in loss of genetic diversity (genetic extinction), which is 17% more likely in vertebrates (69%) than in plants (52%) (Hubbs, 1955; Rhymer & Simberloff, 1996; McCarthy, 2006; Todesco *et al.*, 2016; de Souza *et al.*, 2019). Genetic extinction which results in the loss of distinct species genotypes is attributed to heterospecific mating caused by the lack of reproductive isolation (Rhymer & Simberloff, 1996; Allendorf *et al.*, 2001; Todesco *et al.*, 2016). Incompatible parental regulatory alleles will result in lowered hybrid fitness (Kulmuni & Westram, 2017; Go & Civetta, 2020; McGirr & Martin, 2020).

Allelic incompatibilities within regulatory regions can promote elevated or lowered hybrid gene expression compared to the parental species (transgressive gene expression) (McGirr & Martin, 2020). Since gene expression is maintained by stabilising selection, transgressive expression is predicted to distort exceptionally regulated developmental processes during hybrid development and thus reduce hybrid fitness (Go & Civetta, 2020; McGirr & Martin, 2020). The cellular differentiation pathways for hybrid development were mis-expressed in F1 hybrids of Bahamian and Caribbean *Cyprinodon* pupfishes, acting as a reproductive barrier (McGirr & Martin, 2020).

Hybridisation can also stimulate hybrid fitness (Grant & Grant, 1992; Grant & Grant, 1996; Mallet, 2005; Grant *et al.*, 2003; Grant & Grant, 2010). The fitness of the *Geospiza fortis* (Medium Ground Finch) × *Geospiza scandens* (Common Cactus Finch) hybrids was observed to be greater than the parental species, as comparatively elevated hybridisation generated more environmentally adapted

hybrids that possess an intermediately sized beak linked to efficient nutrient retrieval from opuntia seeds (Grant & Grant, 1996; Mallet, 2005; Grant *et al.*, 2003; Muller, 2008; Grant & Grant, 2010). Finch hybridisation was observed using molecular marker analysis where sixteen *G. fortis* specific microsatellite loci primers reflected increased polymorphisms in *G. scandens* (Grant & Grant, 1996; Petren, 1998; Petren *et al.*, 1999; Mallet, 2005; Grant & Grant, 2010; Grant & Grant, 2019).

1.4.3.7 Monitoring of hybridisation

The identification of phenotypic hybridisation is challenged by the establishment of hybrid-like morphological variants because of mutations or polymorphisms of ancestral origin (Mallet, 2005). However, genomic DNA (gDNA) sequence data may provide a strong indication of introgressive hybridisation, providing an adequate distinction between heterospecific gene flow and ancestral polymorphic content (Mallet, 2005). Adequate distinction between the two phenomena is established through rigorous molecular analysis and recent advancements in genetic studies have led to swift molecular marker development to tackle these phenotypic hybridisation inquiries utilising gDNA (Mallet, 2005; Muller, 2008; Harrison & Larson, 2014; Grant & Grant, 2019). These studies can aid in conservation efforts should results reflect a threat to genetic diversity and integrity (Muller, 2008; Abdurakhmonov, 2016).

Highly polymorphic microsatellites (or short tandem repeats (STRs)) and single nuclear polymorphisms (SNPs) are the most widely utilised markers in ecological studies. STR generational mutation rates range from 10^3 to 10^6 per cell and SNP genetic variation ranges from 1×10^{-4} to 6×10^{-3} per nucleotide (Muller, 2008; Grant & Grant, 2010; Abdurakhmonov, 2016; Vieira *et al.*, 2016; Grant & Grant, 2019; Bravo *et al.*, 2021). These markers are used in avian evolutionary mapping studies (Reed *et al.*, 2003; Lamichhaney *et al.*, 2015), genetic diversity studies (Eo *et al.*, 2011; Dawson *et al.*, 2013) and STR genomic composition studies (Fan *et al.*, 2018). The fundamental motivation for using these markers in hybridisation investigations is the occurrence of a variant genotype in dynamic noncoding regions (Abdurakhmonov, 2016; Vieira *et al.*, 2016; Moran *et al.*, 2021).

Genetic studies were conducted to assess possible hybridisation of Mallards and YBDs in South Africa (Brown *et al.*, 2019; de Souza *et al.*, 2019; Stephens *et al.*, 2019). Brown *et al.* (2019), de Souza *et al.* (2019), and Stephens *et al.* (2019) used mitochondrial and microsatellite markers to

assess the inferred interspecies introgressive hybridisation of the Yellow-billed Duck × Mallard hybrid within parts of South Africa. Brown *et al.* (2019) sequenced 75 YBD mitochondrial DNA (mtDNA) loci using optimised mtDNA control region specific primers on gDNA blood extracts. Sequence data is then compared to published Mallard sequence data (n = 50) to assess Mallard Duck genomic introgression into the YBD genome. The 2019 study found that even though nuclear DNA specific loci showed limited evidence to support species introgressive hybridisation, marker-based evidence from the mtDNA loci indicate ancient species gene flow (Brown *et al.*, 2019). Brown *et al.* (2019), concluded that despite little nuclear evidence to indicate Yellow-billed Duck × Mallard introgressive hybridisation, there is a need to monitor the YBD genome if Mallard populations continue to increase.

Similarly, de Souza *et al.* (2019) used twelve DNA microsatellite loci primers that included five markers established by Maak *et al.* (2003) and two mitochondrial control region specific primers to sequence twenty YBD, twenty Mallard Duck, and eight possible hybrid amplicons produced from gDNA blood extracts. Upon sequence comparison the study found that all inferred phenotypic hybrids possess mitochondrial DNA specific to the Yellow-billed Duck and showed little indication of gene flow between inferred species. Findings also did not confirm the idea that hybridisation between the YBD and Mallard is common in central and north-western South Africa. de Souza *et al.* (2019), concluded that hybridisation could be elevated in regions where Mallard Duck populations are also elevated, thus regular hybridisation monitoring should be conducted throughout South Africa. Stephens *et al.* (2019) conducted a similar experiment with a larger sample size (n = 355, 234 YBDs and 121 Mallards or possible hybrids) and found that the data suggest possible introgressive hybridisation within the Mallard genome. Using the Wilcoxon signed-rank test (Woolson, 2007), data revealed a greater gene flow estimate from YBD populations to Mallard Ducks of approximately 0.0129 migrants per generation (95% credible interval – 0.002 to 0.031, $p < 0.001$, Z = 6073), relative to the 0.0019 migrants per generation (95% credible interval – 0.005 to 0.005) in the inverse direction (Stephens *et al.*, 2019).

1.5 Project rationale and aim

Hybridisation with the Mallard is threatening the genetic integrity of the YBD. The genome sequence of the YBD can serve as a means of understanding YBD specific genetics through the generation of an annotated genome and its phylogenetic relationships in relation to other *Anas*

duck species. Comparative genomic analysis against available Mallard Duck genomes can serve to highlight genetic variation between these duck species, which can ultimately serve to identify species-specific molecular markers for enhanced hybridisation analysis to aid in genetic integrity conservation efforts. This study aimed to sequence, assemble, and annotate the first complete YBD genome and to further conduct comparative genomic analyses with available sequences from other duck species in the genus *Anas*, with specific reference to the Mallard.

Based on this aim, the objectives included:

1. Genomic DNA extraction and concentration analysis from a female, YBD blood sample.
2. Whole genome sequencing (PacBio Sequel IIe sequencing).
3. *De novo* genome assembly using Canu v.2.2.
4. Genomic annotation and completeness assessment through the Modular Open-Source Genome Annotator (MOSGA) pipeline.
5. Comparative genomics and phylogenomic analysis of ducks in the genus *Anas*.

Chapter 2: Sequencing, *de novo* assembly and annotation of the first Yellow-billed Duck (*A. undulata*) genome

2.1 Introduction

A genome is a complete set of genetic information for a single organism, thus serving as a starting point for genomic research (Giani *et al.*, 2020). Since the sequencing of the first full eukaryotic genome, *Saccharomyces cerevisiae* in 1996, whole-genome sequencing has grown more efficient, precise, accurate, and cost-effective using next-generation sequencing (NGS) technologies (Goffeau *et al*., 1996; Shendure *et al.,* 2017). Next-generation sequencing refers to the next step of technological improvement in DNA sequencing, which includes second- and third-generation sequencing methods (Slatko *et al.,* 2018).

Second-generation sequencing (SGS) technologies were designed for low-cost large-scale genome sequencing (Slatko *et al.,* 2018). The SGS technologies outperformed their predecessor, Sanger sequencing, in terms of speed, parallelism, sample multiplexing, sequence coverage, and sensitivity to low-frequency variations (Shendure *et al.,* 2017; Slatko *et al.,* 2018). On the other hand, third-generation sequencing (TGS) methods, are distinguished by real-time single DNA strand sequencing resulting in long read generation relative to SGS technologies without the requirement for pre-amplification (Shendure *et al.,* 2017; Slatko *et al.,* 2018; Giani *et al.*, 2020). Long reads improve assembly quality as reads span larger parts of the genome, allowing for more efficient overlap identification with enhanced low complexity and repeat region resolution (Rhoads & Au, 2015; Shendure *et al.,* 2017; Slatko *et al.,* 2018; Giani *et al.*, 2020).

This study utilised the PacBio Sequel IIe platform (Rhoads & Au, 2015), the forerunner of the PacBio RS II System (Slatko *et al.,* 2018). The PacBio Sequel IIe technology is one of the leading TGS platforms, also referred to as single-molecule, real-time (SMRT) sequencing, which captures sequence information as DNA extracts replicate and produce sequence reads ranging from 30 kb to >50 kb (Rhoads & Au, 2015; Slatko *et al.,* 2018; Logsdon *et al.*, 2020). This eliminates the need for breaks between reading steps and allows it to run faster than SGS platforms (Rhoads & Au, 2015; Slatko *et al.,* 2018).

Avian genomics has taken flight due to their genome compactness, which spans from 0.89 to 2.11 Gb, phenotypic and karyotype diversity (Kapusta & Suh, 2017; Bravo *et al.,* 2021; Cheng *et al.*, 2021; Ksepka, 2021; Waters *et al.,* 2021). To date, 1,162 bird genomes have been sequenced (NCBI Genome Database, 2023). The Bird 10,000 Genomes B10K project has been one of the

largest contributors to bird genome sequencing, with their phase one studies conducted to obtain an evolutionary overview of modern avian species and serving as a starting point for future avian genomic studies (Zhang *et al*., 2014a; Zhang *et al*., 2014b). The project has short-read sequenced 363 individuals spanning 92.4% of all avian families during phase one and two (Feng *et al.* 2020), with phases 3 and 4, targeting genera and subgenera and individual species currently underway.

The sequencing of several bird reference genomes over the last decade has not only enhanced avian phylogeny knowledge but has also enabled genome biology knowledge refinement and improved understanding of the underlying genetic mechanisms that foster phenotypic and functional variation (Zhang *et al*. 2014a, Lamichhaney *et al*. 2015; Bravo *et al*., 2021). The collection of avian genomes, with progressively long-read sequencing data, has also significantly enhanced the understanding of genomic features such as germline-restricted chromosomes and W chromosome genetic content (Bravo *et al*., 2021). Pan-genome avian studies, such as the Zhang *et al*. (2014a) study, show that Aves have fewer genes compared to other tetrapods as a result of high segmental deletions within their genomes. The chicken pan-genome study by Li *et al*. (2022) revealed 1,335 protein coding genes and 3,011 long non-coding genes, the majority of which were core genes previously not identified in mammalian pan-genome studies.

To date, the genomes of two species within the genus *Anas* have been sequenced, namely the Mallard (*A. platyrhynchos* – nine genomes) and Eastern Spot-billed (*A. zonorhyncha*) Duck. This study aimed to sequence, *de novo* assemble and annotate the first genome of the southern YBD subspecies (referred to as the YBD in this study), which is native to South Africa. This will contribute towards our understanding of the genetic and phenotypic diversity associated with the genus *Anas* and concepts underlying Mallard hybridisation and conservation of the species.

2.2 Method

2.2.1 Sampling

Phenotypically pure female YBD blood samples were obtained from a study by Reynolds (2016). These samples were also used to study Yellow-billed Duck × Mallard introgressive hybridisation by de Souza *et al*. (2019). In the initial study the birds were captured using randomised walk-in and mist-netting techniques from the largest southern African waterfowl sanctuary, Barberspan Nature Reserve in the North West Province (Figure 2; 26°35′6.9″ S, 25°35′3.5″ E), between May

and June of 2013 and 2014 (Reynolds, 2016; de Souza *et al.*, 2019; BirdLife International, 2022). Using a capillary tube and a sterile 16-gauge needle, blood was drawn from duck feet and stored at 4ºC in Queen's lysis buffer (de Souza *et al.*, 2019; Stephens *et al.*, 2019). Ethics clearance for capturing and blood extraction was granted by the University of Cape Town Animal Research Ethics Committee (permit number: 2011/V17/GC), and the University of the Witwatersrand Animal Ethics Screening Committee with permit number 2015/04/13/C.

2.2.2 Whole genomic DNA extraction and integrity assessment

Genomic DNA (gDNA) was extracted from *A. undulata* sample YB1665 using a modified protocol for extraction of gDNA from large mammalian blood samples (Mathew, 1984). This method has been shown to work effectively in yielding high molecular weight nucleic acids in a high extraction yield, and further functions as a purification method. The lysis buffer was modified from the Mathew (1984) methodology, with the original protocol utilising a buffer containing 1% Triton X-100, while in this study blood samples were stored and lysed using Queens lysis buffer (QLB) containing 1% N-lauroylsarcosine (Seutin *et al.*, 1991). N-lauroylsarcosine is a strong ionic surfactant used in nucleic acid extractions due to its protein denaturation and protein complex dissociation capabilities (Brown & Audet, 2008; Greco *et al.*, 2014). Seutin *et al.* (1991) developed QLB for long-term avian blood sample ambient temperature storage with no discernible adverse effects induced by repeated thawing and demonstrated that samples stored in QLB yielded elevated concentrations of high-quality DNA. QLB sample storage fostered cellular protein denaturation, while sodium dodecyl sulphate (SDS) utilisation on the pellet post centrifugation facilitated nuclear protein denaturation (Mathew, 1984).

The concentration and purity of the *A. undulata* gDNA was determined using the NANODROP 2000 (ThermoFisher Scientific, USA). Genomic integrity was assessed using standard agarose gel electrophoresis designed by Tiselius (1937), where 5 µl of gDNA was added to a 1.5% agarose gel, which was run at 100V in 1× Tris aminomethane (Tris), Boric acid and EDTA (TBE) buffer (containing equal parts of Tris and boric acid at 89 mM and 2 mM of EDTA). The gDNA sample was sent to Mr DNA (Molecular Research LP, USA) for sequencing on the Pacific Bioscience (PacBio) Sequel IIe sequencing platform. Further sample analysis was conducted with the Qubit® dsDNA HS Assay Kit (Life Technologies-Thermo Fisher Scientific, USA) at Mr DNA to verify

sample concentration and fluorescent electrophoretic analysis was done through E-Gel SizeSelect 2% agarose gel (Invitrogen-ThermoFisher Scientific, USA) to evaluate sample length.

2.2.3 PacBio continuous long read (CLR) and high-fidelity (HiFi) read genome sequencing

2.2.3.1 Library preparation

The PacBio Sequel IIe platform (Rhoads & Au, 2015), was used for monoisolate whole genome continuous long read (CLR) and HiFi sequencing of the extracted gDNA (Slatko *et al.,* 2018). All steps from library construction through to genome sequencing were conducted at Mr DNA (Texas, USA). In brief, library construction began with gDNA shearing using a Covaris G-tube (Covaris Inc., USA), followed by a DNA insert length analysis using the Agilent 2100 Bioanalyzer (Agilent Technologies, USA). An average insert size of 8,207 bp was observed, which fits within the required PacBio insert range of 1 kb to >100 kb (Logsdon *et al.*, 2020). A PacBio compatible gDNA library was created by generating a SMRTbell DNA molecule from ~1 g of gDNA using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA). A SMRTbell is a closed circular DNA molecule composed of a double-stranded DNA insert and two single-stranded pair-ended adapter hairpins ligated on each end of the DNA insert (Figure 2.1a) (Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020). The gDNA library was then subjected to DNA damage and end repair to ensure downstream enzymatic reactions are conducted proficiently, with concurrent adapter barcode ligation to determine which sample a particular read originated from (Ohtsubo *et al.*, 2019; Logsdon *et al.*, 2020). Library DNA concentration and read lengths were determined using the Qubit® dsDNA HS Assay Kit (Life Technologies - Thermo Fisher Scientific, USA), and the Agilent 2100 Bioanalyzer (Agilent Technologies, USA). The final YBD gDNA library had a concentration of 3.05 ng/µl, with an average library length of 8,855 bp.

2.2.3.2 PacBio sequencing

Following library construction, molecules between 6 and 10 kb were selected using a BluePippin Size-Selection System (Sage Science, Inc., USA). The Sequel II Binding Kit 2.0 (Pacific Biosciences, USA) was used to bind prepared libraries to immobilised DNA polymerase within multiple nanophotonic chambers known as SMRT Cells 8M or Zero - Mode Waveguide (ZMW) units (Figure 2.1b) (Eid *et al.*, 2009; Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020).

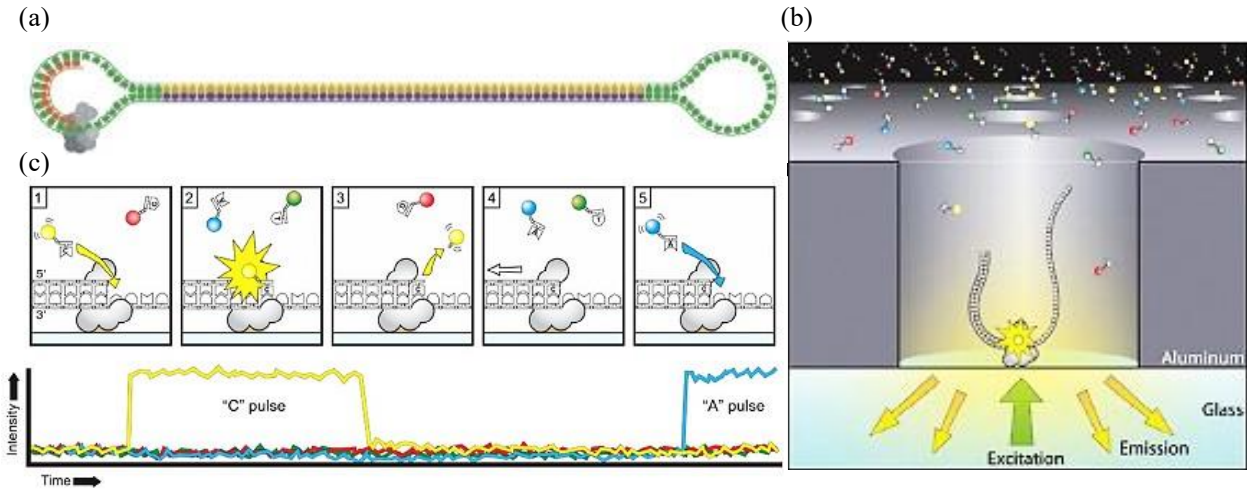


Figure 2.1: PacBio long-read sequencing framework. (a) Immobilised SMRTbell molecule-DNA polymerase complex consisting of the DNA insert (yellow and purple), ligated pair-ended hairpins (green), DNA polymerase in grey and primer in pale orange. (b) SMRT cell/ZMW unit containing the polymerase-DNA complex, which facilitates fluorescently labelled nucleotide (multi-coloured molecules concentrated on top of the cell opening) excitation through controlled light emission. (c) Fluorescence detection sequence within ZMW, with the corresponding depiction of nucleotide specific fluorescence emission intensity over time. (1) Phospho-linked fluorescent cytosine DNA polymerase active site binding (2) results in label excitation and fluorescence emission within the ZMW region. (3) Label dissociation and diffusion away from the ZMW detection zone caused by 3' dNTP phosphodiester bonding. (4) Active site freeing caused by polymerase translocation, (5) resulting in subsequent complementary nucleotide active site binding. Adapted from Rhoads and Au (2015).

The resultant DNA-polymerase -SMRTbell complex drives replication through the correct uptake of four distinct phospholinked fluorescent nucleotides at the enzymatic active site (Eid *et al.*, 2009; Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020). The binding of these fluorophores results in nucleotide excitation which is then detected within the ZMW units (Figure 2.1b) (Eid *et al.*, 2009; Rhoads & Au, 2015). Both CLR and HiFi sequencing were conducted for the YBD genome. PacBio Sequel IIe technology was used to capture sequence data, and raw base calls were exported from the sequencing device and imported into SMRT Link to create HiFi reads using the CCS algorithm (Logsdon *et al.*, 2020).

2.2.3.3 Continuous long read (CLR) and High-fidelity (HiFi) read generation

CLR sequencing utilises DNA libraries containing target DNA inserts greater than 30 kb; however, the platform is compatible with insert sizes ranging from 1 kb to +100 kb (Logsdon *et al.*, 2020). In the case of the YBD genome, CLR sequencing was performed with libraries bigger than the average library size of 8,855 bp, where DNA polymerase makes one or more passes (forming CLR subreads) around the template due to the larger insert size (Rhoads & Au, 2015; Logsdon *et al.*, 2020). In contrast to the most common PacBio sequencing method, High-Fidelity (HiFi) sequencing, is the most recent PacBio sequencing method that generates highly accurate (>99.9%) long (>10 kb) reads (Hon *et al.*, 2020; Logsdon *et al.*, 2020; Sim *et al.*, 2022). These reads are produced through circular consensus sequencing (CCS) of SMRTbell molecules containing shorter DNA inserts, ranging between 10 and 30 kb (Hon *et al.*, 2020; Logsdon *et al.*, 2020). The relatively smaller insert size enables more efficient DNA polymerase action, resulting in the capacity to perform multiple library passes, thus forming an extremely long read composed of multiple subreads that are in the forward and reverse direction (Figure 2.2; Hon *et al.*, 2020; Logsdon *et al.*, 2020). The raw subreads are then aligned and computed using the CCS algorithm (Hon *et al.*, 2020; Logsdon *et al.*, 2020).

CLRs are less accurate, with accuracy ranging from 87 to 92% compared to Illumina short-reads accuracy of >99.9% (Logsdon *et al.*, 2020). Despite the comparatively high error rate, CLRs have two qualities that counterbalance this attribute, namely, generated reads are virtually a Poisson sampling of the target genome, with a uniform and stochastic error distribution (Eid *et al.*, 2009; Myers, 2014). Poisson read distribution suggests that a specific sequence depth (sequence coverage) exists that guarantees no gap formation (Myers, 2014). On the other hand, randomised uniform error distribution suggests that the accuracy of consensus sequence *k* increases with an increase in the number of reads (Myers, 2014, Garrett & Grisham, 2016). This enables downstream assembly, error correction and polishing via tools such as Racon, FALCON and Canu (Chin *et al.*, 2016, Koren *et al.*, 2017, Vaser *et al.*, 2017). To minimize a bulk of stochastic misreads and reach a baseline HiFi read accuracy of 99%, the CCS algorithm often requires 3 to 4 subreads from the same long read (Hon *et al.*, 2020; Logsdon *et al.*, 2020).

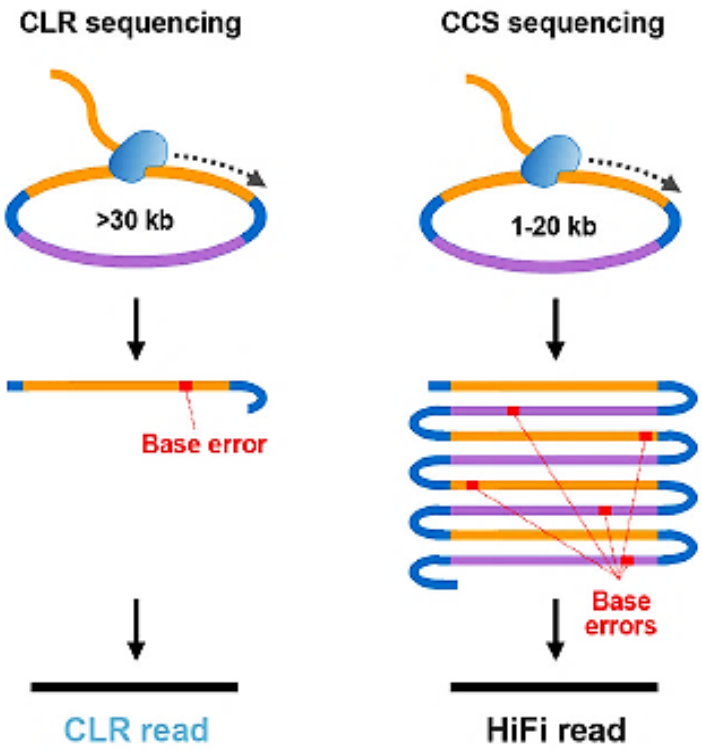


Figure 2.2: PacBio Sequel IIe continuous long read (CLR) and high-fidelity (HiFi) sequencing mechanism. CLRs are typically produced using SMRTbell molecules with DNA inserts larger than 30 kb (forward strand in yellow, adapter molecules in blue, and reverse strand in purple). The larger insert size allows for a single or few DNA polymerase passes, causing the error rate to range from 8 and 15%. Circular consensus sequencing (CCS) generates HiFi reads, which contain DNA inserts ranging from 10 to 30 kb. Smaller DNA inserts facilitate multiple polymerase passes, resulting in multiple subreads. The consensus sequence of these reads is computed to produce a single HiFi read with an error rate ≤1%. Adapted from Logsdon *et al.* (2020).

2.2.3.4 Sequence data quality assessment

The raw CLR and Hifi YBD genome sequence data received from Mr DNA (Texas, USA) and subjected to FastQC v.0.11.9 (Andrews, 2010) to assess data quality. Quality metrics which were assessed included average read lengths, read quality and GC content.

2.2.4 *De novo* long read genome assembly with Canu

2.2.4.1 *De novo* assembler selection

De novo assembly is the process of concatenating read sequences into contigs without mapping onto an existing genome, which avoids the effects of reference-based bias (Chin *et al.,* 2016; Chu *et al.,* 2017; Koren *et al.,* 2017; Vaser *et al.,* 2017; Liao *et al.,* 2019; Hon *et al.,* 2020; Logsdon *et al.,* 2020). *De novo* assembly primarily functions by establishing one of two graphical representations: the overlap layout consensus (OLC) string graph as well as the de Bruijn graph, both depicting interconnected small genomic regions linked via common sequence prefixes and suffixes (Al-Okaily, 2016; Liao *et al.,* 2019; Rizzi *et al.,* 2019). There are two assemblers specifically tailored to reduce assembly bottlenecks by utilising more than one core assembly strategy, generated by PacBio and Oxford Nanopore Technologies (ONT) are Canu and FALCON (Rizzi *et al.,* 2019; Logsdon *et al.,* 2020). The former uses a k-mer-based OLC approach, the latter functions through a hierarchical genome assembly (HGA) approach with OLC string graph integration (Chin *et al.,* 2016; Koren *et al.,* 2017).

Comparative studies of various available assemblers have shown Canu to be the most effective for the assembly of a wide range of eukaryotic genomes, including those of *Caenorhabditis elegans* (nematode) and *Oryzae sativa* (rice) (Vaser *et al.,* 2017, Koren *et al.,* 2017; Lang *et al.,* 2020; Ruan & Li., 2020). The assembler has also been shown to be an effective *de novo* assembler and corrective tool for avian genomes. This includes the recent *Emberiza elegans* (Yellow-Throated Bunting) ONT long read assembly (Hu *et al.,* 2022), the correction of raw PacBio CLRs for 20 chicken genomes (Li *et al.,* 2022), and the CLRs assembly of Z and W - associated contigs of the *Dromaius novaehollandiae* (Emu) (Liu *et al.,* 2021). Furthermore, from its release in 2015 to December 2022, Canu has been cited 4,284 times, and as a result, the raw CLR and Hifi reads were assembled using Canu v.2.2. (Koren *et al.*,2017).

2.2.4.2 Canu: Pipeline architecture

Canu assembles genomic data through an optimised overlap layout consensus (OLC) approach (Lin *et al.,* 2016; Koren *et al.,* 2017; Rizzi *et al.,* 2019). The Canu pipeline assemble reads in three main phases, which can be performed sequentially or individually: error correction to produce

corrected reads; read trimming to identify and remove adapter molecules and chimeric sequences; and finally read assembly to produce an assembly graph and contigs (Koren *et al.*, 2017). The first step in each phase generates an indexed store of input reads, a k-mer histogram, an indexed store of overlaps, and compiles summary statistics (Koren *et al.*, 2017). Each step is initiated by overlap identification between all the pairs of input reads (Koren *et al.*, 2017). While the overlapping technique differs between stages, each overlapping step tallies k-mers within reads, identifies overlaps between the reads, and produces indexed overlaps (Koren *et al.*, 2017).

Step 1: Error correction

Error correction in Canu commences with the MinHash Algorithm Process (MHAP) approach. This algorithm is frequently utilised for rapid and accurate similarity-based overlapping and filtering using MinHash sketches (Sivic & Zisserman, 2003; Chum *et al.*, 2008; Koren *et al.*, 2017; Nikaein & Sharifi-Zarchi, 2022). MinHash sketches are constructed in stages, beginning with decomposing reads into constituent k-mers (Berlin *et al.*, 2015; Nikaein & Sharifi-Zarchi, 2022) and subsequently hashing them into numerical k-mer identifiers (integer fingerprints) (Berlin *et al.*, 2015; Koren *et al.*, 2017). Computationally read sketches require less storage compared to a collection of read k-mers (Berlin *et al.*, 2015; Koren *et al.*, 2017; Nikaein & Sharifi-Zarchi, 2022). MHAP incorporates k-mer term frequency, inverse document frequency weighing (*tf-idf)* into MinHash sketches as a multiplicative combination of the number of times a k-mer occurs inside a read and the overall uniqueness of the k-mer across all reads (Chum *et al.*, 2008; Berlin *et al.*, 2015; Koren *et al.*, 2017). The prevalence of low weighting k-mers within sketches will be reduced, resulting in a more even distribution of indexed k-mers (Chum *et al.*, 2008; Koren *et al.*, 2017).

Step 2: Trimming

Canu utilises the overlap-based trimming (OBT) algorithm (Miller *et al.*, 2008) for corrected read trimming, which involves re-overlapping the corrected reads and expelling unsupported sequences (Koren *et al.*, 2017). Instead of MHAP in the error correction phase, OverlapInCore is used to create a gapped alignment of the corrected reads to precisely identify trim points. Individual reads are trimmed to the largest segment covered by the overlaps with at least the expected coverage (Koren *et al.*, 2017). Following corrected read trimming, a second pass is performed to detect

residual untrimmed adapter molecules, which are identified when multiple reads share a small common sequence flanked by forward and reverse overlap (Eid *et al.*, 2009; Koren *et al.*, 2017).

Step3: Assembly

Additional filtering steps are conducted because Canu employs the Bogart algorithm, which is an altered Miller *et al*. (2008) strategy for assembly graph construction that works by selecting the best overlaps to create the best overlap graph (BOG) (Koren *et al.*, 2017). Bogart first assembles unitigs using the greedy approach (Al-Okaily, 2016), as described by Miller *et al*. (2008), because long reads typically have high sequence coverage, which can result in more read collisions for reads with the same prefix sequence (Koren *et al.*, 2017). Bogart excludes contained reads, which are reads that are enclosed within other reads (Fasulo *et al.*, 2002). Bogart then randomly selects a k-mer from a specific read for extension until all the best overlaps are exhausted, generating a unitig, which is subsequently removed from the data set, and the process is repeated (Miller *et al.*, 2008; Al-Okaily, 2016; Koren *et al.*, 2017).

Canu then uses a modified version of a PBDAG-Con-based algorithm to generate a consensus sequence for each unitig to form the final contigs (Chin *et al.*, 2013; Garrett & Grisham, 2016; Koren *et al.*, 2017). The consensus sequence is created by concatenating reads from approximate positions based on the best overlap layout consensus (OLC) path within a DAG (Chin *et al.*, 2013; Chin *et al.*, 2016; Koren *et al.*, 2017). Although the consensus sequence generated from the reads is highly accurate because of multiple filtering steps, contigs may still contain indels from inaccurate overlapping. To correct for this all the reads in the unitig are aligned to the best overlaps and added to the DAG which then calls the final contigs (Chin *et al.*, 2013; Koren *et al.*, 2017).

2.2.4.3 Canu assembly of the Yellow-billed Duck genome

The raw CLR and Hifi reads were assembled using Canu v.2.2. (Koren *et al.*, 2017). To optimise genome assembly and to evaluate the efficacy of the different Canu assembly steps, different combinations of the CLR and Hifi data were used in combination (Figure 2.3).

The raw CLR reads were first corrected and assembled (step 1 + 3, bypassing the trimming step). Subsequently, the effect of trimming was evaluated by passing the raw CLR reads through error correction, trimming and assembly (step 1 +2 + 3). Next the corrected (step 1 only) and trimmed (step 1 + 2) read datasets were combined with a Hifi read dataset and assembled with Canu step 3.

Correction and trimming of the Hifi dataset on its own did not yield change in the Hifi read dataset. Finally, CLR data parsed through Canu step 1 +2 were combined with 2x Hifi datasets. The envisaged genome coverage for the CLR data alone, CLR + 1 Hifi dataset and CLR + 2 Hifi datasets are predicted to be 14×, 12× and 57×, respectively.

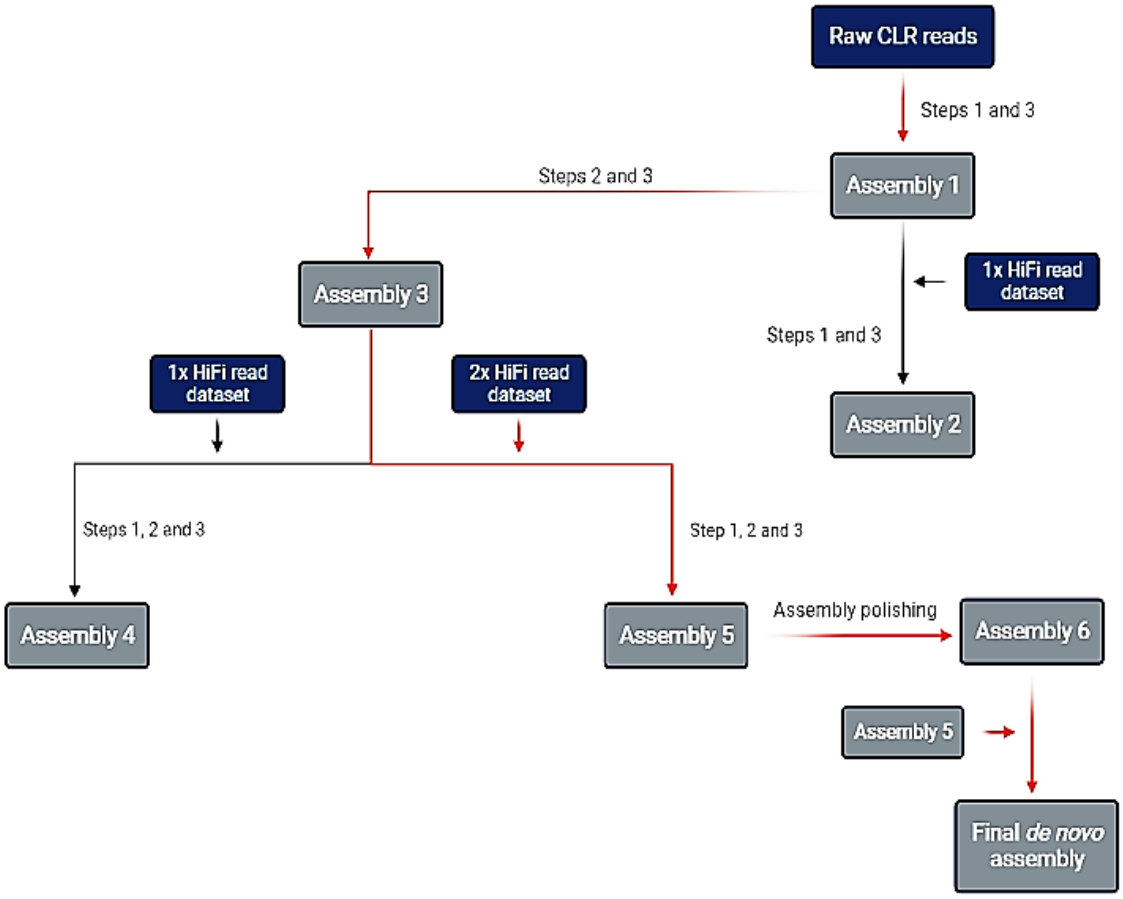


Figure 2.3: *De novo* long-read assembly framework. Grey blocks indicate assembly datasets, and the red line indicates the path to the final *de novo* assembly. Created with BioRender.com.

2.2.4.4 Assembly polishing: Racon

The optimal Canu assembly output (corrected and trimmed CLR + 2x Hifi datasets) was further polished using Racon v.1.4.20. (Vaser *et al.*, 2017). Prior to tool utilisation, the raw CLR read (uncorrected and untrimmed) dataset, in the form of a FASTQ file comprising raw reads and quality information, was first aligned against the Canu contig assembly using minimap2 (Li, 2018). Racon first filters the CLR read-contig overlaps, removing overlaps with a high error rate, followed by rapid edit-distance-based alignment of the filtered overlaps (Vaser *et al.*, 2017). The initial

alignment separates reads into blocks that fall into certain non-overlapping windows on the contig sequence; these read blocks are subsequently screened to remove low-quality blocks (Vaser *et al.*, 2017). A partial order alignment (POA) algorithm then calls the consensus of each window independently, creating a POA graph (Lee *et al.*, 2002), and the final corrected sequence is created by concatenating individual window consensuses together (Vaser *et al.*, 2017). Unpolished contigs were retained in our analyses and were added to the polished assemblies for genome assembly completeness.

2.2.5 Yellow-billed Duck genome annotation

Genomic annotation is the identification of genomic segments, their location and function within a genome sequence (Logsdon *et al.*, 2020; Martin *et al.*, 2021a; Martin *et al.*, 2021b). The final YBD genome assembly was structurally and functionally annotated using the Modular Open-Source Genome Annotator (MOSGA) pipeline (Figure 2.4; Martin *et al.*, 2021a; Martin *et al.*, 2021b).

2.2.5.1 Adapter contamination analysis

The presence of two pair-end ligated adapter molecules of known sequence within the PacBio DNA library (SMRTbell) aids in primer binding for insert sequencing facilitated by polymerase amplification (Rhoads & Au, 2015; Slatko *et al.*, 2018; Logsdon *et al.*, 2020). The adapter molecules generated within a single SMRT cell are then recognised and cleaved, and adapter contamination analysis is used to evaluate the efficacy of this step (Rhoads & Au, 2015). Possible adapter hairpin contamination from the SMRTbell molecule was then assessed using the NCBI tool VecScreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/).

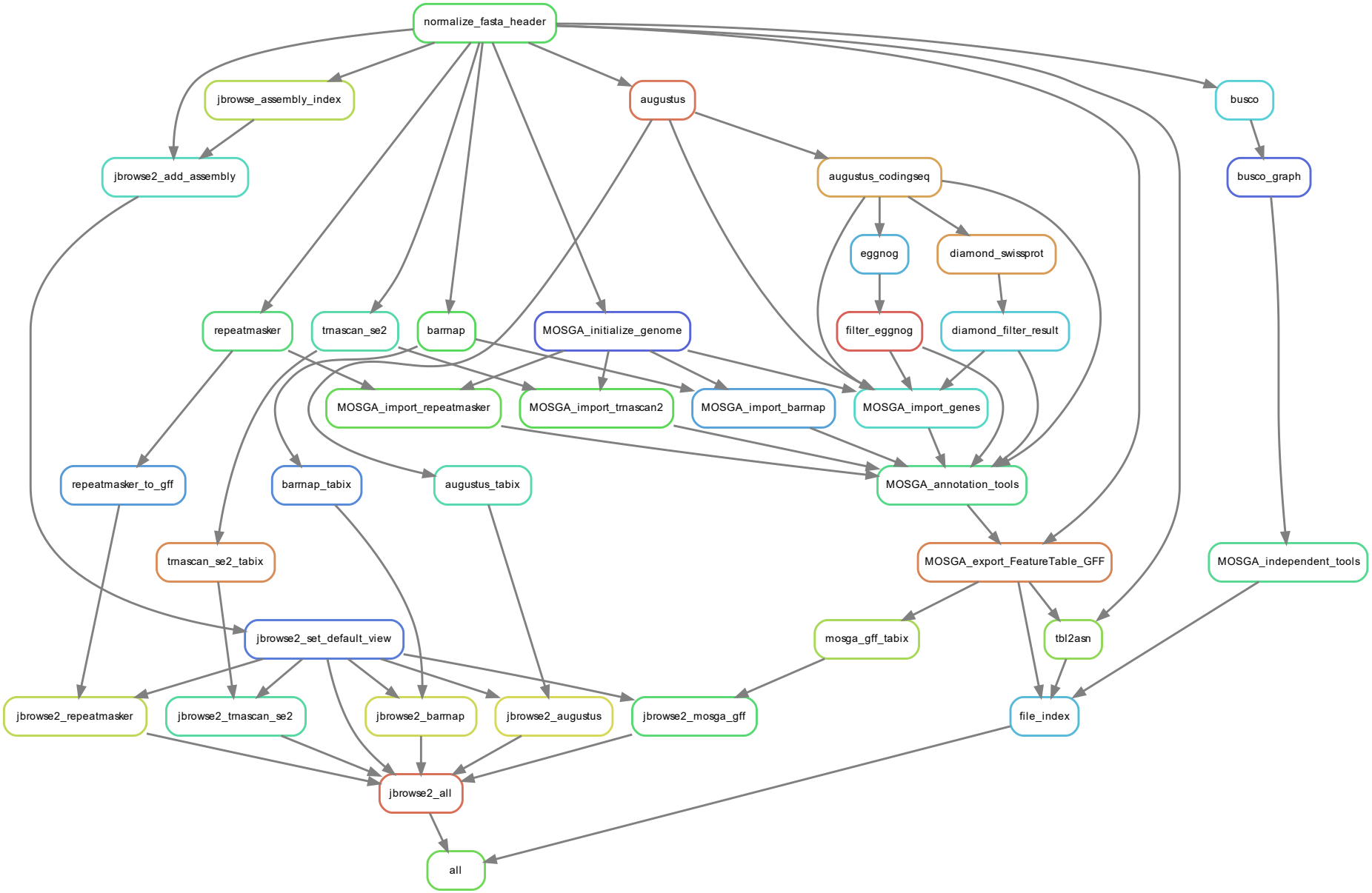


Figure 2.4: MOSGA pipeline sequential software database utilisation and subsequent file outputs for the Yellow-billed Duck genome. MOSGA generated image (Martin *et al.*, 2021a; Martin *et al.*, 2021b).

The tool identifies foreign DNA sequences within the query sequence through BLAST similarity cross-referencing against the UniVec database utilising parameters necessary for adapter contamination identification (NCBI, 2022).

2.2.5.2 Genome completeness analysis

Genome completeness was assessed using the BUSCO v.4.1.4 (Simão *et al.*, 2015) implementation in MOSGA (Martin *et al.*, 2021a; Martin *et al.*, 2021b). This tool was further used by MOSGA during the early stages of the pipeline to refine the gene model and improve annotation with AUGUSTUS (Stanke & Waack, 2003; Stanke *et al*., 2008), where the normalised data is used to assess assembly completeness with reference to expected single-copy gene content. The fact that evolutionary related organisms possess conserved single orthologs drives tool functionality (Simão *et al.*, 2015). BUSCO employs HMMER v.3.3.2 (Wheeler & Eddy, 2013) to search benchmarked clade-specific ortholog data sets for orthologs in the annotated gene dataset, in this case the AUGUSTUS gene predictions (Simão *et al.*, 2015). The OrthoDB-stored Aves_odb10 lineage data set, which has an evolutionary distance of 4.0 relative to the query sequence and contains 8,338 Aves-specific orthologs, was used in this study.

2.2.5.3 Repeat element identification and annotation

Most eukaryotic genomes contain repeat genomic elements, which are sequence segments that are repeated hundreds or thousands of times in the genome (Biscotti *et al.*, 2015). However, the concentration of repetitive nucleotide sequences interspersed throughout avian genomes is lower compared to most eukaryotic genomes, measured between 2 and 10% except for the *Picoides pubescens* (Downy Woodpecker) at 22.15% (Zhang *et al.*, 2014a; Zhang *et al.*, 2014b; Weissensteiner & Suh, 2019). The identification of these repetitive genomic regions can aid in conservation studies through species-specific microsatellite design (Brown *et al.*, 2019; de Souza *et al*., 2019; Stephens *et al.*, 2019). Repeat elements were predicted in the YBD genome using the MOSGA implementation of RepeatMasker v.4.1.1 (Smit *et al.*, 2013). This tool employs HMMER v.3.3.2 to perform nucleotide homology searches against a database, in this case Dfam v.3.2 (Hubley *et al.*, 2016). Dfam v.3.2 is an open-source database that contains 6,953 different DNA-based repeat families, each of which is characterised by a multiple sequence alignment and a profile HMM (Hubley *et al.*, 2016). The HMM model was trained using the *Anas platyrhynchos*

(Mallard Duck; NCBI ID: 8839) dataset, to adjust search parameters to be specific to the Yellow-billed Duck genome for increased accuracy (Wheeler & Eddy, 2013).

2.2.5.4 *Ab initio* gene prediction

Genes were predicted on the final YBD genome assembly using AUGUSTUS v.3.4.0 (Stanke & Waack, 2003; Stanke *et al*., 2008). To improve gene prediction MOSGA uses BUSCO v.4.1.4 (Simão *et al.*, 2015) in the early stages of the pipeline to assess assembly completeness (Martin *et al*., 2021a; Martin *et al*., 2021b). The pipeline then uses the entire single copy ortholog hit dataset to retrain AUGUSTUS and further refine the model parameters for improved gene prediction accuracy (Martin *et al*., 2021a; Martin *et al*., 2021b). This algorithm uses a generalised hidden Markov model (GHMM) or a semi-Markov conditional random field (semi-CRF) model to predict protein-coding genes and their structure for a specific sequence (Stanke & Waack, 2003; Stanke *et al*., 2006; Stanke *et al*., 2008; Hoff & Stanke, 2019; Scalzitti *et al*., 2020).

2.2.5.5 Yellow-billed Duck functional annotation

For functional annotation of the AUGUSTUS-predicted protein datasets, the proteins were aligned against two independent databases, eggNOG 5.0 (Huerta-Cepas *et al*., 2019) and SwissProt (Bairoch *et al*., 2004). The SwissProt alignment was performed in MOSGA using DIAMOND (Buchfink *et al*., 2015) with the following parameters: E-value of 0.001, sequence identity > 30, and query coverage of 50. Alignment against the eggNOG database was performed with eggNOG-mapper v.2.0.6 (Huerta-Cepas *et al*., 2017) with an E-value of 0.001, sequence identity ≥ 0, and query coverage ≥ 0. eggNOG 5.0 orthology database is a curated database that spans the entire Tree of Life (5,090 organisms), providing extensive annotation coverage while minimizing redundancy (Huerta-Cepas *et al*., 2019; Cantalapiedra *et al*., 2021). The eggNOG database 5.0 includes Gene Ontology Consortium terminology (2015), the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps (Kanehisa *et al*., 2014), and Clusters of Orthologous Groups of proteins (COGs) functional classification (Galperin *et al*., 2015).

Gene Ontology (GO) describes and categorizes genes based on the expression product and denotes the function, relative cellular location, and the biological process in which it is involved (Gene Ontology, 2015). KEGG is an integrated database composed of fifteen databases to facilitate high-

level understanding of biological functionality and the usefulness of biological systems based on nucleotide and amino acid level information (Kanehisa *et al.*, 2014). The database produces metabolic pathway maps that illustrate the metabolic pathways query sequences are involved in (Kanehisa *et al.*, 2014). The COGs database employs a family-based approach, in which the function of a known member of the protein family is used to assign function to the entire protein family or COG and describe potential functionality range when there are multiple members (Galperin *et al.*, 2015).

2.2.5.6 RNA prediction and annotation

Transfer RNAs (tRNAs) were predicted with the MOSGA implementation of tRNAscan-SE v.2.0.8 (Chan *et al.*, 2021). tRNAscan-SE employs covariance models to obtain amino acid sequence and secondary structural tRNA data for complete tRNA gene exploration (Eddy, 1992; Eddy & Durbin, 1994; Nawrocki & Eddy, 2013; Chan *et al.*, 2021). Covariance model construction is achieved through Infernal 1.1 implementation, and the tool then uses the models to rapidly search for new hits within sequence databases via sequence alignments (Nawrocki & Eddy, 2013). The normalised assembly data was run on tRNAscan-SE v.2.0.8 (Chan *et al.*, 2021), with default parameters; search mode as eukaryotic, an infernal first pass with a cut-off score of 10 (Nawrocki & Eddy, 2013), and allowing for an isotype-specific scan.

Ribosomal rRNAs were predicted using the MOSGA implementation of Barrnap (Basic Rapid Ribosomal RNA Predictor). Barrnap predicts ribosomal RNA genes using HMMER 3.1 (Wheeler & Eddy, 2013), which uses rRNA-gene-specific HMM models built from full-length seed alignments from the RefSeq (O'Leary *et al.*, 2016), Rfam (Griffiths-Jones *et al.*, 2003) and SILVA databases (Quast *et al.*, 2013) to perform sequence homolog searches within databases and sequence alignments (Seemann, 2022). Barrnap was run with the following parameters: eukaryotic as the chosen kingdom to indicate which database to use, lencutoff at 0.8 (reject 0.25), which is the minimum percentage of the full rRNA element length for partial matches, where matches < 0.25 of the expected length are excluded, and the minimum E-value at 0.00001.

2.3 Results and Discussion

2.3.1 Yellow-billed Duck genome raw data quality assessment

The gDNA of sample YB1665 had a final concentration of 240 ng/µl, with purity ratios of 1.75 (A260/A280) and 2.03 (A260/A230). These readings indicate a relatively pure sample according to Koetsier & Cantor (2019). E-Gel SizeSelect 2% agarose gel (Invitrogen-ThermoFisher Scientific, USA) revealed the sample length measured at >15Kb. PacBio sequencing of the libraries yielded both a CLR and Hifi dataset. The CLR dataset comprised of 2,245,741 reads, with an average read length at 9,999 bp (range: 11 bp – 366,502 bp) (Figure 2.5a). The HiFi dataset comprised of 1,091,314 reads, with an average length at 6,999 bp (Figure 2.5b; range: 80 – 28,431 bp).

FastQC analysis of the CLR and Hifi datasets showed high Phred scores for the Hifi data (92), while two Phred score peaks of ~11 and 92 were observed for the CLR data (Figure 2.6). The Phred score (Q) is a statistical indicator of the likelihood of proper nucleotide incorporation within a particular read during sequencing (Ewing & Green, 1998; Prosdocimi *et al.,* 2003; Sathyanarayanan *et al.,* 2018). A Q score of 20 to 40 indicates good data quality, which corresponds to sequence correctness of 99 to 99.99% and error probability of 1 to 0.01% (Ewing & Green, 1998; Prosdocimi *et al.,* 2003; Sathyanarayanan *et al.,* 2018). Values >40 imply a higher possibility of accurate nucleotide incorporation, with sequence accuracy greater than 99.99% and error possibilities less than 0.01%, whereas Q scores <20 indicate poor data quality, with a higher error rate and lower accuracy (Sathyanarayanan *et al.,* 2018). HiFi reads had a greater and more consistent Q score, culminating at 92 (Figure 2.6b), compared to the two CLR peaks at just under 11 and a lower peak at 92 (Figure 2.6a). The occurrence of two peaks suggests that the CLR data is a combination of low- and high-quality reads. The low-quality reads are at a higher concentration indicated by the higher peak, implying that the CLR base call accuracy is between 90 and >99.9999%, with most CLRs generated having a 10% error rate. FastQC however, did not detect any low-quality reads in either sequence data set.

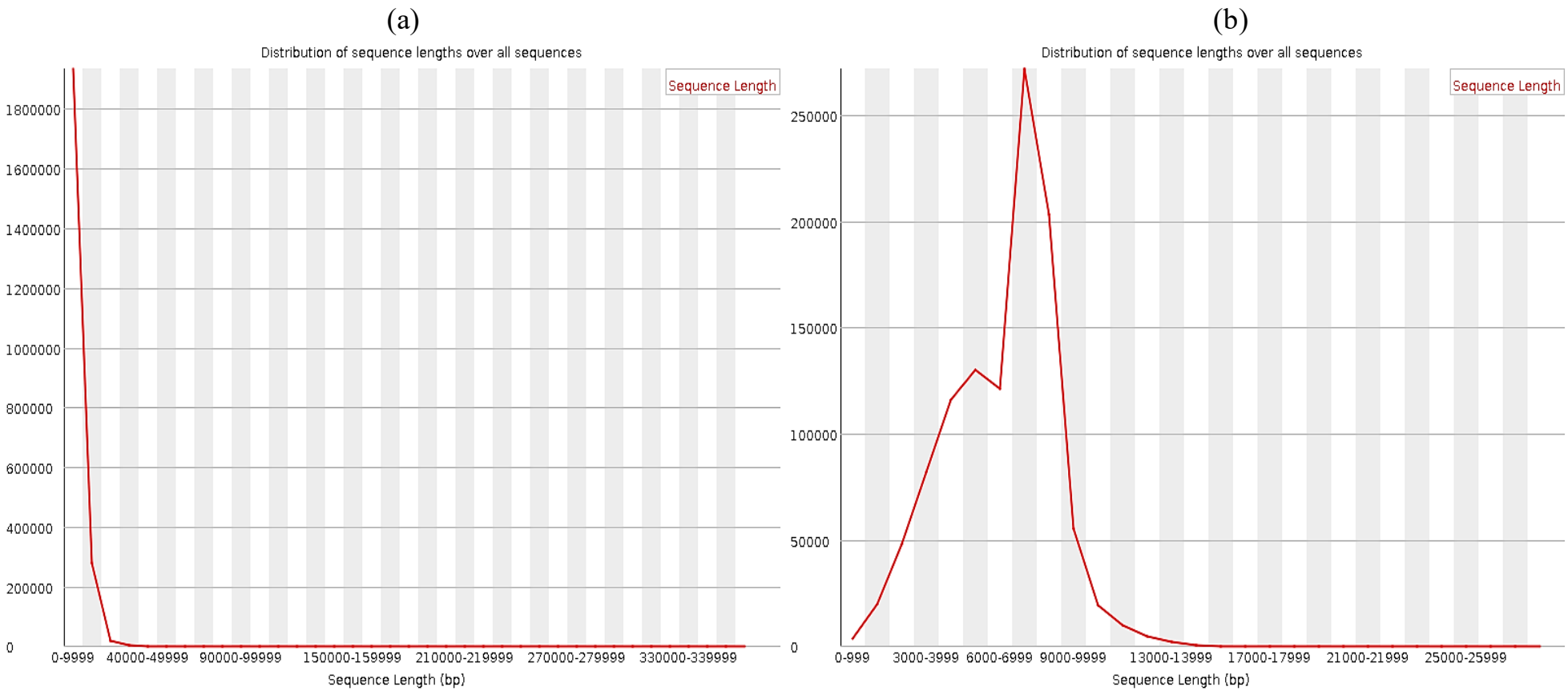


Figure 2.5: CLR and HiFi read length distribution. (a) displaying CLRs with an average read length of ~10 kb and (b) displaying HiFi read length distribution at ~7 kb.

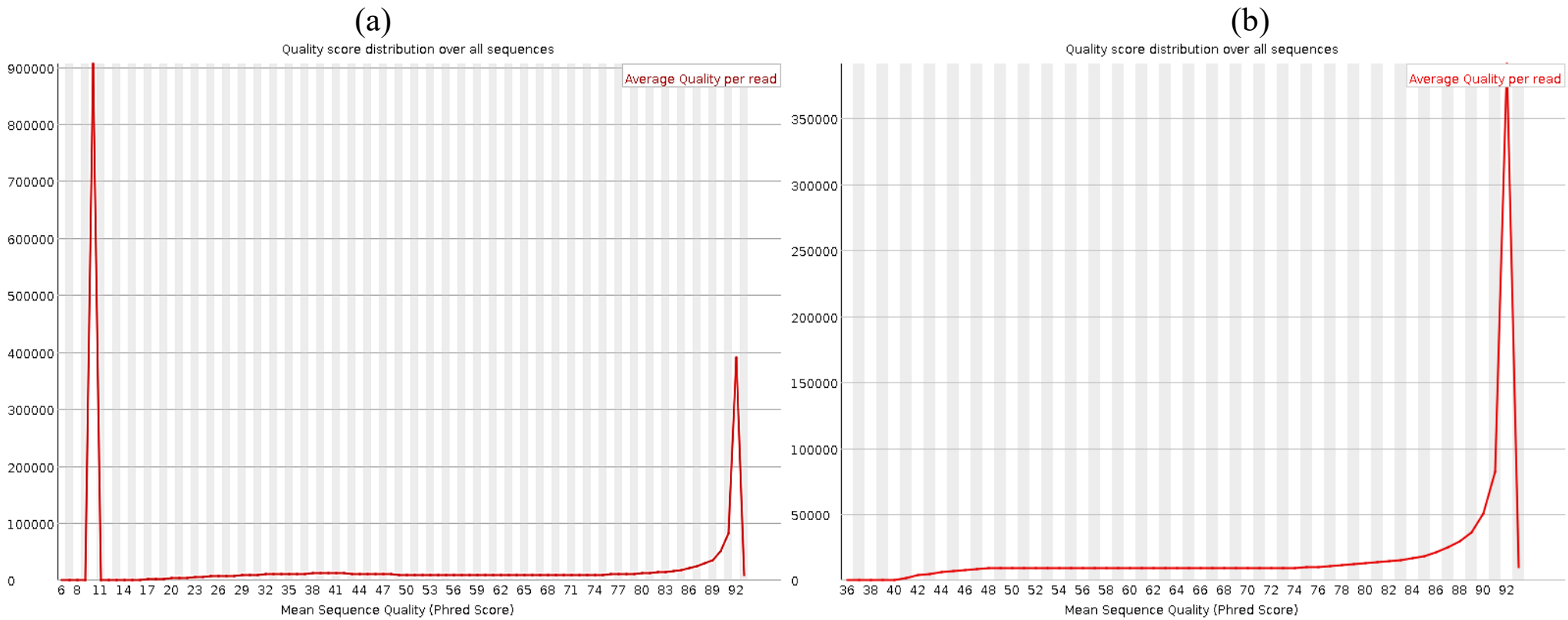


Figure 2.6: Quality score (Q) read distribution. (a) representing CLRs data and (b) representing HiFi read data. CLRs have two peaks at 92 and the highest at <11, suggesting that most CLRs have sequence accuracy of 90%. The Q score for the HiFi reads peaks at 92, indicating sequence accuracy >99.9999%.

The CLR raw sequence data had an expected average GC content of 42%, which is comparable to the genomes of other *Anas* species (Table 3.2). The average GC content for the HiFi reads, however was slightly lower at 41%. The minor CLR GC content elevation could be attributable to adapter contamination, which was also seen by Sim *et al*. (2022), who found adapter molecule contaminants with marginally higher GC percentages in 53 of the 55 HiFi reads. Because sequences from next-generation sequencing (NGS) platforms with extremely GC-rich and/or GC-poor genomic regions have nonuniform or no read coverage across the genome, GC content distribution is also a metric of sequence quality, assessing read coverage (Chen *et al.,* 2013; Sathyanarayanan *et al.*, 2018). As a result, GC-balanced genomic regions are preferentially selected, a phenomenon known as GC bias (Benjamini & Speed, 2012; Chen *et al.,* 2013; Beauclair *et al.,* 2019). The GC bias is assessed using a plot of the number of reads versus the GC percent per read, where a reference distribution (theoretical distribution in Figures 2.7a and b) is generated by computing the modal GC percent using the observed data to generate a unimodal distribution. Both sequence data sets follow a similar GC content distribution, with both peaking at the theoretical peaking point of 39%, but at a read frequency above the theoretical distribution, which indicates possible GC bias (in blue on Figures 2.7a and b).

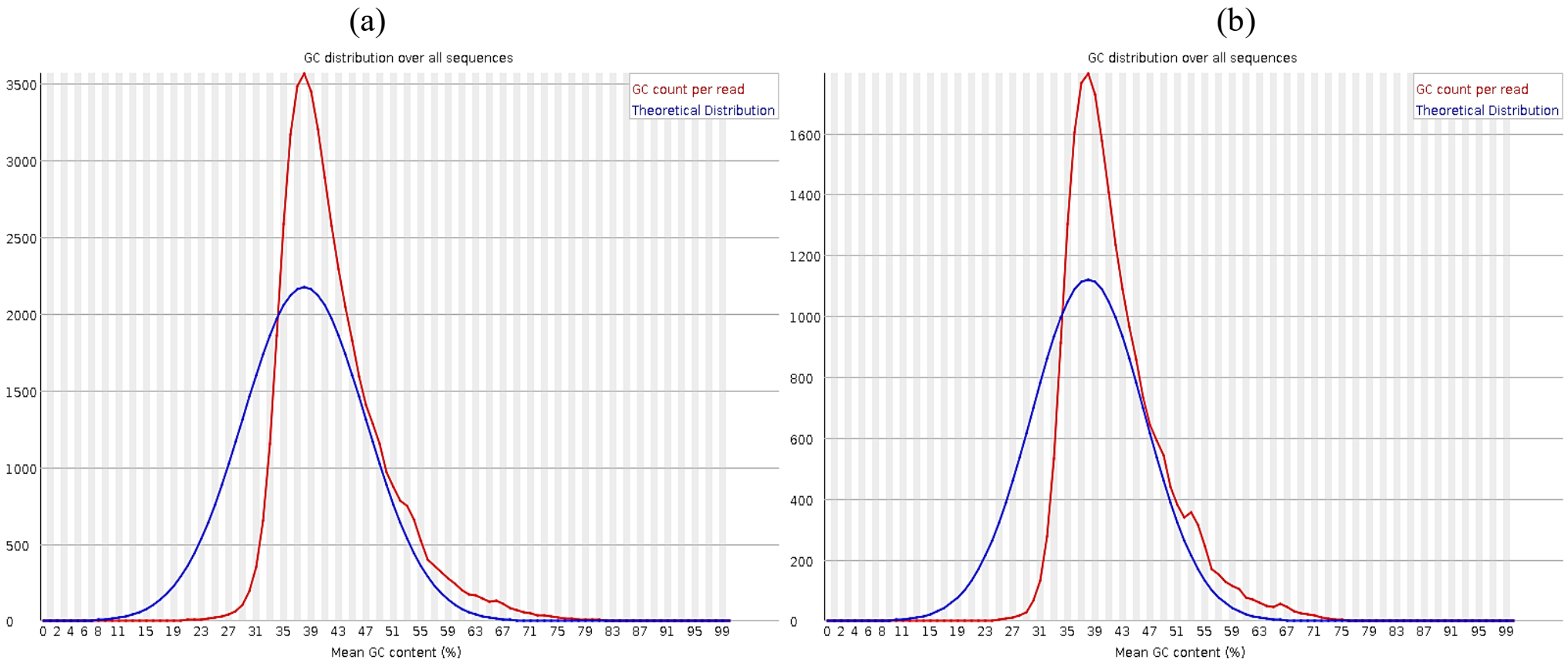


Figure 2.7: GC content distribution across total CLR and HiFi reads. CLRs (a) and HiFi reads (b) both peaking at 39%, but above the unimodal theoretical distribution, suggesting GC bias for both data sets.

2.3.2 *De novo* assembly

De novo assemblies were undertaken using different combinations of the CLR and Hifi datasets and steps of the Canu assembler (Koren *et al*.,2017). The assembly metric for all datasets are compared in Table 2.1. Assembly of the CLR data with Canu with the trimming step bypassed (Assembly 1) yielded the least contiguous assembly (23,558 contigs; N50: 64,776; 1.2 Gb genome). In the second assembly (Assembly 2), a 1x Hifi dataset was added to the CLR data and Canu was run again with the trimming step omitted. This yielded an assembly with substantially lower number of contigs and an N50 value 2.03× greater than Assembly 1, but a much smaller genome, at 978 Mb. In Assembly 3 the CLR data was solely Canu assembled but with the trimming step (Step 2) included. Assembly 3 resulted in a >2× decrease in contig number and minimum contig length, while the maximum contig length had a >6 increase compared to Assembly 1 (trimming step excluded). Furthermore, the third assembly is more contiguous relative to assemblies one and two, despite it having a shorter assembly length compared to the first assembly.

Table 2.1: *De novo* assembly metrics assembled using Canu v.2.2

Assembly No.	Data assembled	Canu steps	No. of Contigs	Min contig length (bp)	Max contig length (bp)	N50 (bp)	Total assembly length (bp)
1	CLR	1 + 3	23558	3811	487947	64776	1120512875
2	CLR + 1x Hifi	1 + 3	15019	1927	857165	131497	978124728
3	CLR	1 + 2 + 3	11268	1723	3121120	242534	1094223644
4	CLR + 1x Hifi	1 + 2 + 3	10264	3306	3011885	329563	1130603964
5	Assembly 3 + 2×HiFi	1 + 2 + 3	10767	4074	2507718	350863	1159852075
6	Racon polishing Assembly 5	N/A	10469	5319	2510228	353324	1157373542
7	**Assembly 6 + unpolished contigs**	**N/A**	**10767**	**4074**	**2510228**	**351454**	**1161236934**

Bold assembly metrics indicate the final assembly

Since the addition of a single HiFi read dataset, trimming and reassembly improved the contiguity of the first assembly, Assembly 4 incorporated a full Canu run with the CLR data and 1x HiFi

dataset. This assembly comprised 1,004 fewer contigs (10,264 contigs) compared to Assembly 3 (excluding Hifi data), had a longer minimum contig length (3,306 bp) and a shorter maximum contig (3,011,885 bp) compared to the third assembly. To increase coverage Assembly 5 was generated using the CLR data with two HiFi read datasets. This resulted in a 6.07% increase in N50 value and a 2.52% increase in assembly length (1,159,852,075) compared to Assembly 4 with a single Hifi dataset. The decrease in minimal contig length with the addition of Hifi data to the assembly is likely due to the lower HiFi read length.

Racon polishing (Assembly 6) of the final full Canu assembly of the CLR dataset and two HiFi datasets increased this minimal contig length and also increased the contiguity of the assembly, as indicated by the highest N50 value achieved throughout (N50 = 353,324). Those contigs which were not polished by Racon were subsequently added to Assembly 6 to yield the final assembly. Assembly 7 comprises a total length of 1,161,236,934 bp in 10,767 contigs, with an N50 value of 351,454 bp, and a GC content of 42.03%. There are 888 contigs in the final assembly that are larger than the N50 value, with four contigs being >2 Mb.

2.3.3 Analysis of adapter contamination

Vecscreen identified 1,926 adapter contamination matches in 658 contigs (Table 2.2), most of which constituted weak terminal and internal matches. The high number of weak matches could be a result of a fragmented alignment from a single long stretch of DNA, as UniVec contains contiguous sequences ≤ 50 bp in length, which explains the overall hit base pair average of 33.53 bp (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/#Limits). Strong matches make up only 2.54% (49) of the total adapter matches, while weak make up majority of the matches at 90.5% (1,743) which are low confidence matches which often occur through chance. Overall, with 0.047% adapter contamination this is unlikely to have an impact on genome assembly and quality.

Table 2.2: VecScreen data output summary

Type of match	No. of matches	Average match length (bp)	Total match length (bp)	Genomic composition (%)
Weak terminal	697	16.52	11513	0.000991443
Weak internal	1046	34.17	35737	0.003077494
Moderate terminal	23	19.09	439	3.78045E-05
Moderate internal	111	33.23	3689	0.000317678
Strong terminal	10	40.3	403	3.47044E-05
Strong internal	39	57.85	2256	0.000194276
Total/Average	1926	33.53*	54037	0.0046534

* Overall average

2.3.4 Genome completeness analysis

BUSCO analysis of the YBD genome suggests a genome completeness of 93.6% (Figure 2.8). Of the 8,338 BUSCOs, 92.60% are present on the YBD genome as complete single copy orthologs (SCOs), 0.96% are duplicated, 2.58% are fragmented and 3.86% of the Aves BUSCOs are missing from the YBD genome. The *G. gallus* assembly (GCA_016699485), on the other hand, has a completeness of 99.0%, with 98.6% complete SCOs, 0.4% duplicated, 0.2% fragmented, and 0.8% missing Aves BUSCOs (NCBI, 2022).

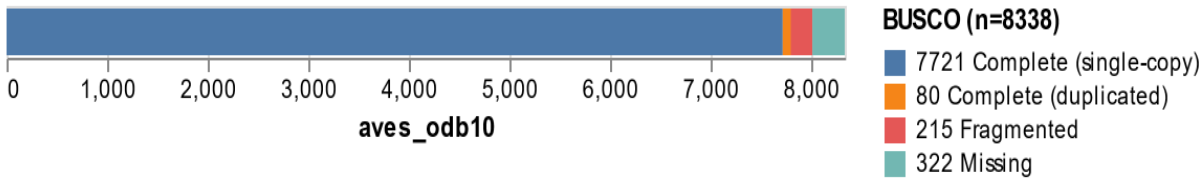


Figure 2.8: BUSCO genome completeness assessment.

2.3.5 Repeat element identification and annotation

A total of 185,221 repeat elements were identified in the YBD genome. The total size of these repeat elements is 118,154,391 bp, or 10.17% of the genome (Table 2.3). This fits the avian repeat element range of ~4 to 15% reported by Zhang *et al.* (2014a, 2014b) The most common repeat class identified is the Long-Interspersed Nuclear Element (LINE) class, which accounts for 15.13% of the repeat elements and 5.68% of the total genome. This is consistent with the trend reported by Bravo *et al.* (2021) in their review of 500 avian genomes spanning 89% of the

recognised avian families at the time of publication. Furthermore, the study identified CR1s as the most common LINEs in avian genomes which is also true for the YBD, with CR1s accounting for 5.67% of the genome (Bravo *et al.*, 2021). RepeatMasker identified 207,160 retroelements and 104,406 DNA transposons.

Table 2.3: Yellow-billed Duck repeat element genomic composition

Repeat element	No. of elements	Genomic length occupied (bp)	Genomic composition (%)
SINEs	47496	5880710	0.51
LINEs	157024	65950389	5.68
LTRs	2640	284363	0.02
DNA transposons	104406	11255698	0.97
Rolling circles	466	49791	0.0043
Unclassified	54529	4034969	0.35
Small RNA	74276	5983134	0.52
Satellites	2	18651	0.0016
Simple repeats	518751	22271380	1.92
Low complexity	78585	4060658	0.35
Total	1038175	119789743	10.3259

2.3.6 RNA prediction and annotation

The tRNAscan-SE v.2.0.8 tool predicted 424 tRNAs on the YBD genome. Of these, 385 decode the twenty standard amino acid codons, one decodes the selenocysteine codon, three have unknown codon annotations, three have isotype mismatches, and 32 are pseudogenes (Chan et al., 2021). The tRNA data is imported into the JBrowse2 accumulator (Martin et al., 2021a; Martin et al., 2021b). Within the accumulator, tRNAs with <70 infernal scores were filtered out, resulting in the filtering of 164 tRNA isotypes. Further data processing was conducted on the filtered 260 isotypes to categorise hits according to the amino acid the tRNA molecules transport (Figure 2.9). The amino acids glycine (11.15%) and serine (11.92%) had the highest tRNA representation (Figure 2.9, in green), while tryptophan, arginine, and methionine only contributed 1.15%, 1.54%, and 2.31% of the total tRNAs, respectively, with no genes transporting histidine or asparagine.

The number of YBD tRNAs is comparable to that observed in the chicken genome, with the latter carrying tRNAs coding for 280 standard amino acids, a single selenocysteine codon, a single unexpected match, a single isotype mismatch, and nine pseudogenes (Chan & Lowe, 2016). However, there is a gene distribution variation, with the highest genes coding alanine (9.25%), and like the YBD, serine has the second highest number of gene hits (7.83%), while 2.14%, 6.41%, and 4.98% encode tryptophan, arginine, and methionine, respectively (Chan & Lowe, 2016). Unlike the YBD, the chicken genome contained genes coding for histidine (2.49%) and asparagine (3.20%).

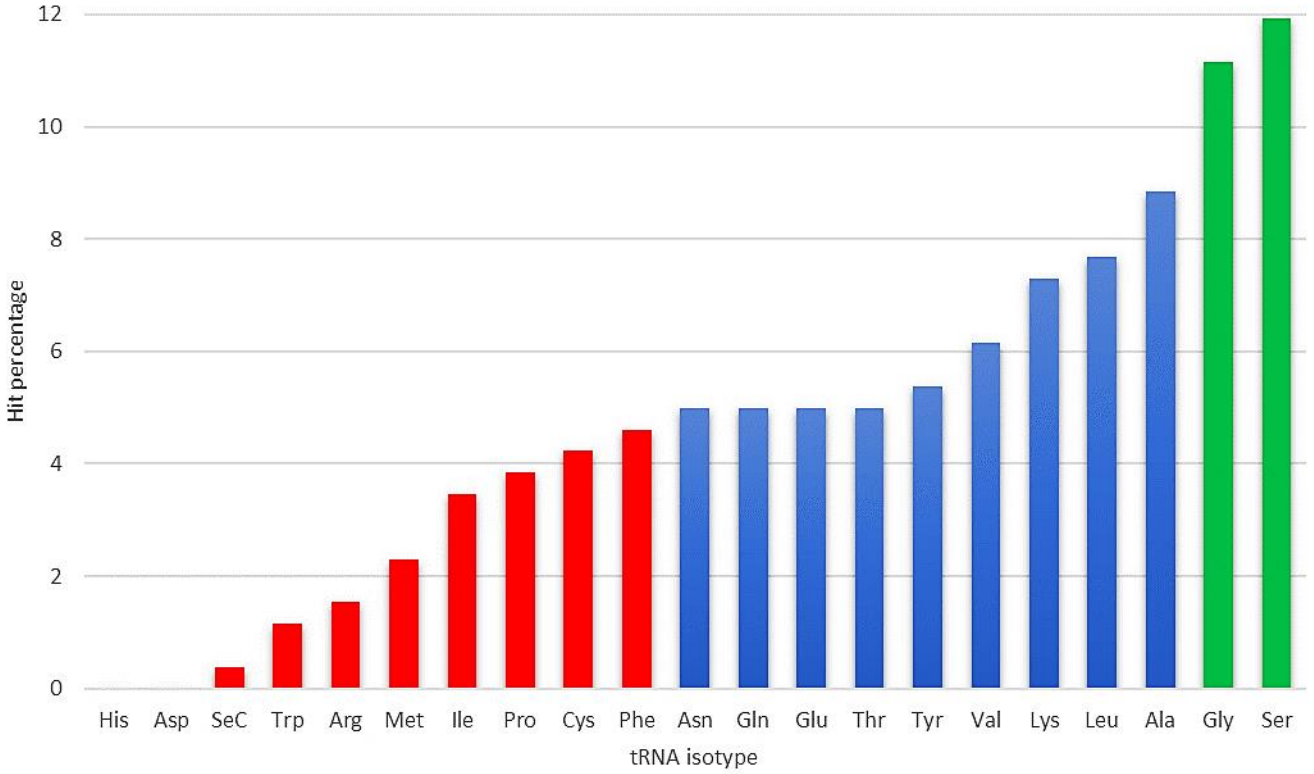


Figure 2.9: Distribution of transfer RNA (tRNA) genomic hits from 260 tRNA isotypes predicted on the YBD genome. Data obtained using tRNAscan-SE v.2.0.8 (Chan *et al.*, 2021).

A total of 91 rRNA matches were predicted on the YBD genome using Barrnap (Seemann, 2022). A total of fifteen 28S rRNA (the large rRNA subunit) elements were predicted, with an average size of 4,060 bp. The 18S rRNA (the small rRNA subunit) and 5.8S rRNA elements each had thirteen matches, with average lengths of 1,801.23 bp and 151.15 bp, respectively. Finally, 39 5S rRNA elements were predicted with an average length of 113.23 bp. The average rRNA element lengths correlate with those of the chicken reported by Dyomin *et al.* (2016), with the 28S rRNA,

18S rRNA, and 5.8S rRNA elements of the latter measuring 4,441bp, 1,823 bp, and 157 bp, respectively. The 5S rRNA subunit, as part of the large ribosomal subunit, promotes protein synthesis by providing structural stability, with the average gene length in this study correlating to Szymanski *et al*. (2002), who measured the element at ~120 bp.

2.3.7 *Ab initio* gene prediction output

Ab initio gene prediction with AUGUSTUS in MOSGA predicted 42,414 unfiltered gene hits on ~78% of the total 10,767 contigs. Of the contigs, nine had ≥ 50 gene hits, with the largest contig (2.5 Mb) containing 63 genes with an average gene length of 19,409 bp. The CDS value is greater than the number of gene hits and this is because the coding sequence is interspersed between intronic regions (Furuno *et al.*, 2003).

Table 2.4: AUGUSTUS output summary

AUGUSTUS output statistic	Unfiltered No. of Hits: Post MOSGA retraining [a]	Filtered No. of Hits: Post accumulator processing [b]
Genes	42414	39258
Average gene length (bp)	10687.52	9840.47
CDSs	217900	189531
Average CDS length (bp)	229.17	238.71
Introns	175486	-
Average intron length (bp)	2296.56	-
Transcripts	42414	39258
Average transcript length (bp)	10687.52	6678.39
No. of Contigs with hits	8370	8093

[a] Unfiltered gene prediction hits pre-accumulator

[b] Filtered gene prediction hits post accumulator comparison with functional annotation outputs

The 42,414 AUGUSTUS gene hits were loaded into the accumulator and compared to the functional annotations detailed in subsequent sections, resulting in data filtering to 39,258 gene hits with an average gene length of 9,840.47 bp (Table 2.4). The accumulator reads all cumulative data from each selected tool, merges and filters the data through comparison, performs quality

assessments, and finally deciphers the appropriate output following Snakemake activation (Martin *et al*., 2021b; Mölder *et al*., 2021).

2.3.8 Functional annotation

A total of 36,046 gene products (84.99% of 42,414 predicted genes) were functionally annotated using eggNOG-mapper (Figure 2.10). The largest proportion of proteins 26.76% (of 36,046 gene products) are involved in "carbohydrate transport and metabolism" (G), followed by 17.16% proteins of "unknown function" (S), and 13.83% of proteins involved in "signal transduction mechanism" (T). No proteins were assigned to the functional categories mobilome: "prophages, transposons" (X) and "general function prediction" only (R). A DIAMOND BLASTx (Buchfink *et al.,* 2015) search on the SwissProt database (Bairoch *et al.,* 2004), identified 17,230 functional proteins (40.62% of the 42,414 predicted genes) (Figure 2.11). SwissProt has relatively fewer functional annotations as it is a more species-specific database (Bairoch *et al.,* 2004). More than 99% of the sequence data stored within the database stems from nucleotide translation of sequences stored within EMBL (Kanz, *et al.,* 2005), GenBank (Benson *et al.,* 2012) and DNA Data Bank of Japan (DDBJ) (Bairoch *et al.,* 2004; Mashima *et al.,* 2016; Cantalapiedra *et al.,* 2021). Despite the utilisation of translated nucleotide sequences from multiple databases, SwissProt is non-redundant, and this is achieved by taking the consensus sequence of the same gene product from multiple organisms (Bairoch *et al.,* 2004). Over 50% of the proteins on the YBD genome annotated with SwissProt are involved in "carbohydrate transport and metabolism" (G) at 67.46%, followed by proteins involved in "inorganic ion transport and metabolism" (P) at 17.60% and "post translational modification, protein turnover, chaperones" (O) at 7.23%.

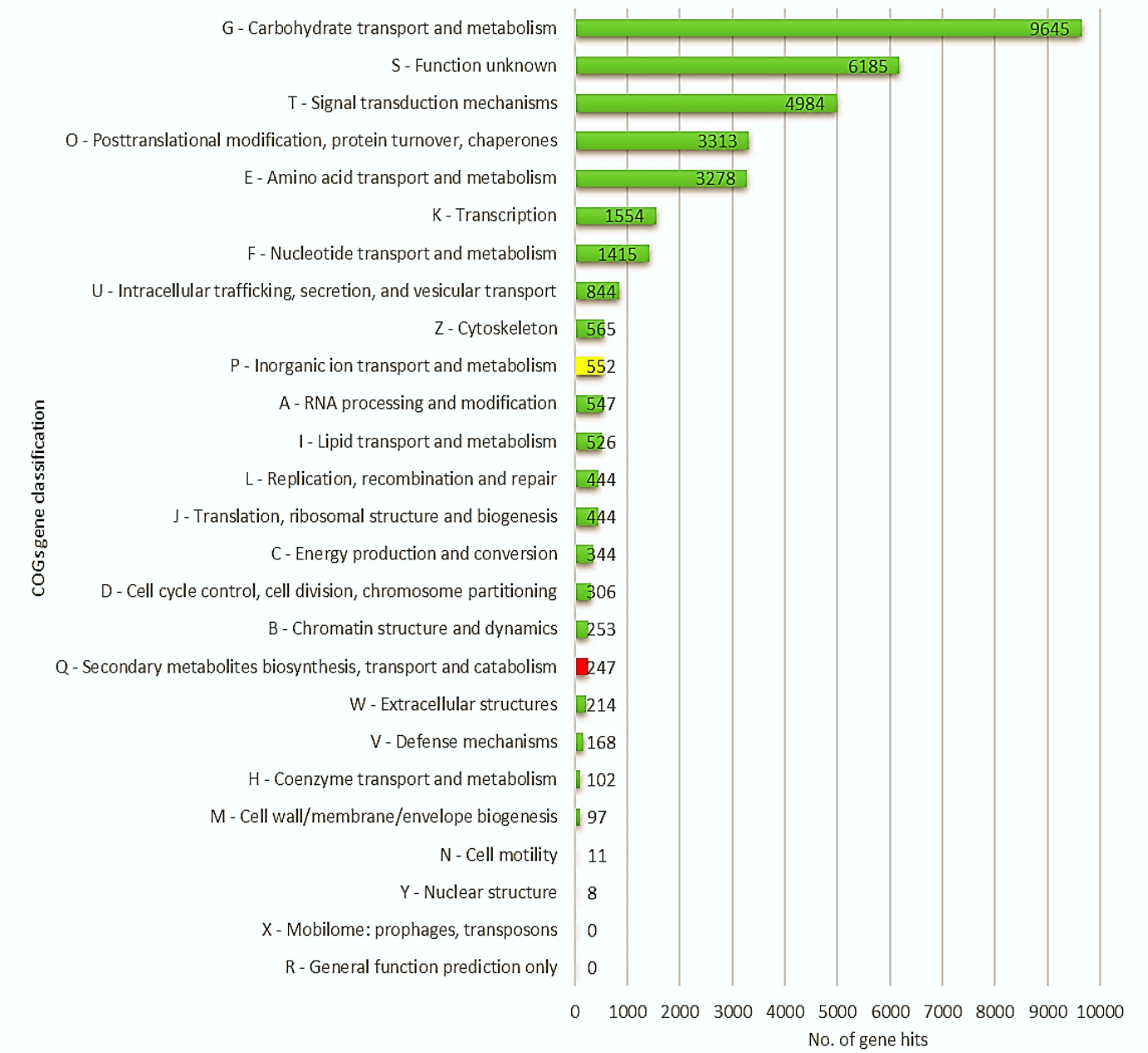


Figure 2.10: Classification of Yellow-billed Duck gene products into the 26 COG functional categories according to eggNOG-mapper alignment against the eggNOG 5.0 database.

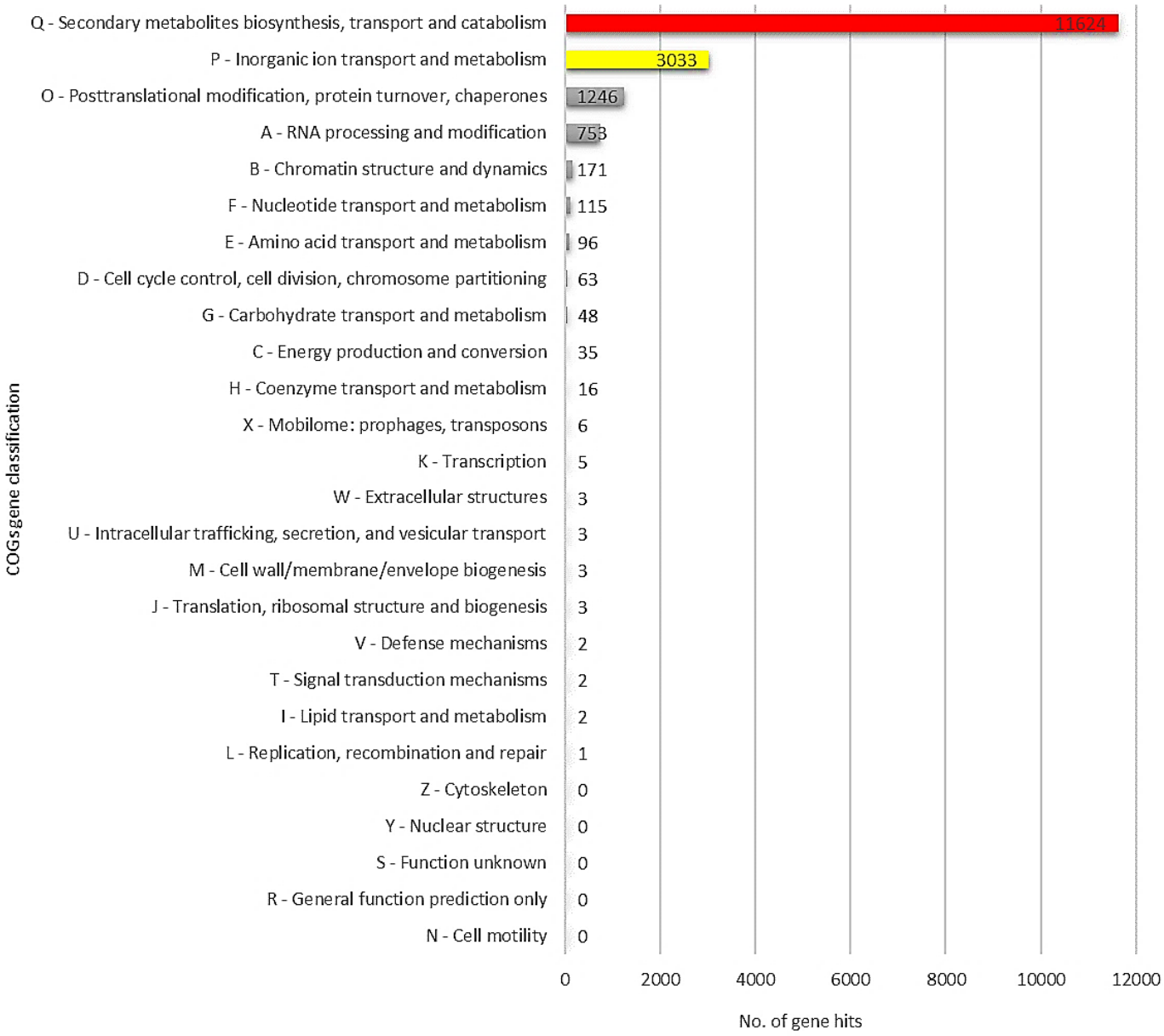


Figure 2.11: Classification of Yellow-billed Duck gene products into the 26 COG functional categories according to DIAMOND alignment against the SwissProt database. In total, 67.46% of the 17,230 functional hits are proteins involved in the biosynthesis, transport, and metabolism of carbohydrates, while fifteen of the twenty-six categories have <100 hits, and five have no gene product hits.

Comparison of the YBD gene set against the SwissProt database (Bairoch *et al.,* 2004) with DIAMOND (Buchfink *et al.*, 2015), resulted in the functional annotation of 17,230 gene products (40.62% of the 42,414 predicted genes). This is substantially (44.37%) lower than annotation against the eggNOG database. However, all the gene hits are proteins assigned to COG categories. In this case, the majority of the proteins (~67.5%) are involved in the biosynthesis, transport, and metabolism (catabolism) of secondary metabolites biosynthesis (Q), with only 0.69% of proteins involved in this function in the eggNOG annotation (Figures 2.10 and 2.11 in red). The second largest COG category (17.60%) obtained from the SwissProt alignment, proteins involved in the transport and metabolism of inorganic ions (P), follows a similar trend with only 1.53% of proteins involved in this function in the eggNOG annotation (yellow in Figures 2.10 and 2.11). The reason for the extensive difference in COG category prediction between the eggNOG and SwissProt annotations remains unclear.

2.3.8.1 Enzymatic annotation

A total of 4,503 (12.49% of 36,046) proteins annotated with eggNOG-mapper were assigned to enzyme classes using the KEGG database (Figure 2.12; Kanehisa *et al.,* 2014; Cuesta *et al*., 2015; McDonald and Tipton, 2021). Transferases, hydrolases, and oxidoreductases make up 92% of all enzymatic hits at 47%, 35% and 10% respectively, which are the three enzymatic classes with the most enzymatic subclasses as defined by Cuesta *et al*. (2015).

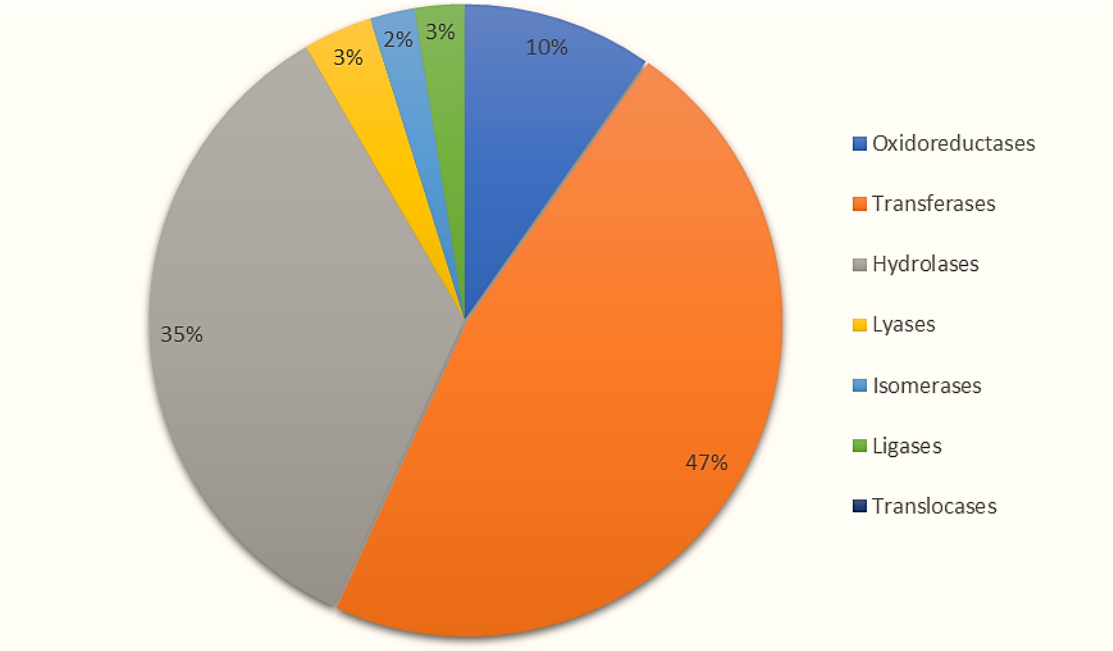


Figure 2.12: The Yellow-billed Duck's enzymatic profile from 4,503 KEGG hits. Transferases and hydrolases make up 82%, with 2,119 and 1,567 enzyme hits, respectively. This is followed by oxidoreductases with 439 hits, while lyases, ligases and isomerases all have <200 hits (160, 115 and 103 hits, respectively).

2.4 Conclusion

The genome of a female *Anas undulata* from Barberspan was sequenced using SMRT sequencing technologies, which produced CLR and HiFi reads. The sequencing datasets were *de novo* assembled to a high-quality draft level by numerous rounds of Canu long-read assembly using different genome assembly parameters, then polished using Racon. The final genome sequence comprises 10,767 contigs with an overall size of 1.16 Mb, an N50 value of 351,454, and genome completeness of 93.6%. Finally, structural, and functional annotation was performed on the *A. undulata* genome using the MOSGA platform and its associated open-source modules. In Chapter 3, this fully annotated genome serves as the basis of comparative genomic approaches to evaluate its relation to other *Anas* species, their phylogenomic relationships and evolution, and to highlight the annotated YBD genome as a resource for the expansion of genus- and species-level avian genome data.

Chapter 3: Comparative genomics of the Yellow-billed Duck

3.1 Introduction

The genome of an organism provides primary insights into its genetic potential, the potential molecular determinants underlying its biological functions and the proteins it produces (Hardison, 2003; Galperin & Koonin, 2010; Brown *et al.*, 2019; de Souza *et al.*, 2019; Bravo *et al.*, 2021). However, the link between genotype and phenotype remains largely obscure for the majority of genes and the phenotypic characteristics for which they code (Nussinov *et al.*, 2019). With the advent of fast, inexpensive and accurate sequencing technologies, it has become easier to sequence the genomes of multiple organisms displaying distinct genotypic and phenotypic characteristics (Shendure *et al.*, 2017; Slatko *et al.*, 2018; Bravo *et al.*, 2021). This has led to a revolution in the development of tools to compare the genomes of these different organisms and elucidate the molecular basis underlying their distinct characteristics and hence better understand the links between genotype and phenotype. In the last decade, the avian NCBI reference genome library has increased from two in 2012 to 596 genomes as of August 2022 (NCBI, 2022). Of these genomes only two duck species (*A. platyrhynchos* and *A. zonorhyncha*) belong to the *Anas* genus, which accounts 6.25% (of 32 species) of the genus (Gill *et al.*, 2021; NCBI; 2022).

Simply stating species' genomic content does not provide a complete picture of how the genotype results in phenotypic behavioural traits, how the species interacts with its external environment, or it relatedness; thus, comparative genomics and the establishment of a phylogenetic framework is required (Hardison, 2003). In comparative genomics, the carefully selected avian genomes are compared using genomic parameters such as repetitive elemental composition, functional protein content, sequence homology, and other bioinformatic based parameters (Hardison, 2003; Zhang *et al*., 2014a). Phylogenomics assesses species relatedness at various taxonomic levels by using specific molecular markers such as evolutionary conserved mitochondrial markers (Castresana, 2000; Delsuc *et al*., 2005; Zhang *et al*., 2014b).

Interestingly, despite elevated avian diversity, with over ten thousand recorded species and 19,883 subspecies (Gill *et al.*, 2021), ~3.10% of all avian species (339) have been sequenced and annotated (NCBI, 2022). Avian evolutionary links remain incomplete because phylogenetic studies such as those by Zhang *et al*. (2014a; 2014b) and Feng *et al*. (2020), publications from the Bird 10,000 Genomes (B10K) project, have made studying ties between key lineages such as orders or families their focal point. However, despite these knowledge gaps, Aves are still one of

the most well-studied clades in comparative genomics, with major avian genomic contributors like B10K shifting towards genus and species-level refinement (Genome 10K Community of Scientists, 2009; Bravo *et al*., 2021).

In Chapter 2, the genome of *A. undulata*, the Yellow-billed Duck (YBD) was sequenced, assembled, and annotated. This duck is closely related to the Mallard Duck, with the latter duck interbreeding with *A. undulata*, placing the YBD under threat. The genomes of nine Mallard Ducks, including two wild, six domesticated ducks and a single wild and domestic cross, have been sequenced. Furthermore, the genome of the wild Eastern-spotted billed ducks has also been sequenced. Here, by means of phylogenomics, the evolutionary relationships between the YBD, Mallard Duck (six individuals) and Eastern Spot-billed Duck will be analysed. Subsequently, the presence/absence of protein coding genes between the genomes were analysed, as well as an evaluation of the copy of number of key genes among the different bird genomes. This was to elucidate genes which are overrepresented in the YBD or its counterpart duck relatives. Finally, through codon analysis of single-copy orthologues conserved among the compared ducks, unique YBD proteins compared to its relative, to elucidate differences in the selective pressures in the genes that code for them were identified. This chapter aims to expand past the YBD genome and comparatively analyse it with available *Anas* genomes retrieved from NCBI.

3.1.1 *Anas* species to avian genomic analysis

The Zhang *et al*. (2014a and 2014b) studies are a combination of comparative genomics and phylogenomic studies by first sequencing, assembling, annotating, and comparing forty-eight genomes from 36 different bird orders. The studies include a monoisolate *A. platyrhynchos domestica* bird genome sequence (assembly: GCA_000355885) and at the time of publication all orders within the Neognathae clade and two of the five Palaeognathae orders (NCBI currently has six *Palaeognathae* orders). This resulted in genome lengths spanning 1.04 and 1.26 Gb (Zhang *et al*., 2014b). Fan *et al*. (2018), on the other hand, employed comparative genomics to analyse STR genomic content from twenty-three ancestral Mallard Ducks (wild *A. platyrhynchos*) and twenty-six *A. platyrhynchos domesticus* (Pekin Duck) to better understand waterfowl STR genomic content. The evolutionary analysis highlighted that genes with divergent STRs have an increased potential of causing phenotypic alterations during the domestication of the ducks (Fan *et al*., 2018).

The study of the common domestic duck, which is mostly considered to be the *A. platyrhynchos domestica*, in research such as Serjeantson (2014), Zhang *et al.* (2018), and Wang *et al.* (2020), has also contributed to the advancement of bird domestication knowledge. The Serjeantson (2014) study is a duck domestication review that laid the groundwork for studies like the Zhang *et al.* (2018) and Wang *et al.* (2020) studies, by outlining the lack of duck domestication knowledge. Zhang *et al.* (2018) and Wang *et al.* (2020) investigated phenotypic selection and duck domestication chronology using whole-genome resequencing on wild Mallards and domesticated ducks, followed by comparative genomics. The Zhang's *et al.* (2018) comparative genomics component found that genes involved in brain and neural development have been subjected to significant positive selection during domestication (Zhang *et al.,* 2018). Wang *et al.* (2020) found that genes involved in lipid metabolism were strongly selected during domestication, while the MITF gene is linked to white plumage expression.

3.2 Methods

3.2.1 Assembly comparative analysis

The genome assemblies of nine Mallard Ducks (*Anas platyrhynchos*) and one Eastern Spot-billed Duck (*Anas zonorhyncha*) were obtained from the NCBI Genome Assembly database (Table A1). The reference genome (GCA_016699485) of *Gallus gallus* (domestic chicken) was included as an outgroup. The genomes of all comparative taxa were annotated using the Modular Open-Source Genome Annotator (MOSGA) (Martin *et al.*, 2021a; Martin *et al.*, 2021b) pipeline as per the Yellow-billed Duck (Chapter 2 section 2.2.5). MOSGA assesses assembly completeness with BUSCO v.4.1.4 (Simão *et al.,* 2015); thus, BUSCO outputs for each comparative assembly were analysed to determine individual assembly completeness. For each comparative genome, assembly metric such as the number of contigs, the minimum and maximum contig lengths, the N50 value, ungapped assembly length and GC content were also examined (Table 3.1). The genomes of three mallard taxa, namely Wild×Dom (GCA_900411745) and Pekin-Female (GCA_000355885) and *A. platyrynchos* CAU-Wild (GCA_008746955) did not meet the quality criteria and were removed from further analyses (Tables 3.1 and 3.2).

3.2.2 Annotation comparative analysis

MOSGA genome annotations and parameters such as gene number, mean gene length, functional annotations, RNA annotations and repeat element profiles from the nine comparative taxa were evaluated.

3.2.3 Orthology inference

The protein datasets derived from the MOSGA implementation of AUGUSTUS for all comparative genomes were subjected to OrthoFinder v.2.4.1. (Emms & Kelly, 2015) to identify orthologous proteins, i.e., those shared between different combinations of the bird comparative genomes or unique to the genome of a specific bird. Orthologue identification with OrthoFinder was done using default parameters with DIAMOND (Buchfink *et al.,* 2015).

3.2.4 *Anas* core genome phylogeny

Single copy orthologues (SCOs) i.e., those shared among all compared genomes (all eight *Anas* genomes and the chicken outgroup) as only one copy per genome were identified from the OrthoFinder output. These were individually aligned using the mCoffee implementation of T-Coffee v.13.45 (Notredame *et al.,* 2000). All the individual alignments were then concatenated, and poorly aligned blocks were removed with GBlocks v.0.91 (Castresana, 2000). A maximum likelihood phylogeny (Figure 3.3) constructed using IQ-Tree2 (Minh *et al.,* 2020) with the optimal evolutionary model as predicted using MODELTEST (Posada & Crandall, 1998) and ultrafast bootstrapping (n = 1,000 replicates). To further refine the *Anas* phylogeny, a second unrooted VT+F+I+G4 tree was built without the outgroup depicted in Figure A1.

3.2.5 Pangenome analysis

The pangenome is the complete collection of genes within a taxon, such as a species or a genus. It is comprised of the core genome comprising genes shared by all members of the taxon and an accessory genome (Morneau, 2021). The pangenome captures all genetic diversity within a taxon and decreases genetic analysis bias caused by utilising a single reference genome (Wang *et al.,* 2023). The pangenome of the genus *Anas* was determined using the protein complements of the available eight *Anas* genomes. From the OrthoFinder output, the core genome (those proteins shared among all compared genomes), accessory genome (those shared by some but not all

compared genomes) and unique (only found in a single genome) fractions were determined (Figure 3.5). The Eggnog and SwissProt outputs from MOSGA were used to functionally annotate the core, accessory and unique genome fractions.

3.2.6 Paralogue analysis

Paralogous genes are those genes that are derived from gene duplication events since the last common ancestor of the compared organisms (Gogarten & Olendzenski, 1999; Koonin, 2005). Thus, differences in the number of paralogues for a specific organism are indicative of gene duplication subsequent to speciation (Magadum *et al.,* 2013). Gene duplication underpins evolution, by providing the foundation for the acquisition of novel genetic material on which mutation, drift, and selection can occur, resulting in novel gene functions (Magadum *et al.,* 2013). Genes displaying extensive paralogy in specific *Anas* ducks, were determined using the OrthoFinder output (section 3.2.3). The gene copy number for each orthogroup for each comparative genome was first determined. Orthogroups with a cumulative (across the eight compared duck genomes) paralogue sum ≥50 were selected, yielding 180 orthogroups. The ducks were then clustered according to species or origin (wild vs domesticated ducks) and compared. Orthogroups were sorted according to the standard deviation of each species-specific gene copy number, and the top fifty from each respective comparison were selected. The functions of the high copy number paralogues were determined using eggNOG-Mapper (Huerta-Cepas *et al.,* 2017). YBD paralogues were compared to those on the genomes of all other *Anas* ducks, and the overall gene copy levels of wild ducks were compared to domestic birds.

3.2.7 Codon diversity identification using BaCoCa

The curated (with Gblocks) individual alignments of each SCO protein conserved among the high-quality genomes were analysed using BaCoCa v.1.105 (Kück & Struck, 2014). BaCoCa statistical analysis was conducted for rapid codon diversity identification in SCOs of specific ducks or groups of ducks (species or wild vs domesticated) to indicate diversification. BaCoCa is a base composition calculator that employs several statistical analytic methodologies to identify potential reference bias from alignments, which could alter phylogenetic tree construction (Kück & Struck, 2014). The tool produced a tab-delimited file in which the Chi-square Test of Homogeneity and relative composition frequency variability (RCFV) values were used to evaluate sequence homogeneity and taxon specific compositional bias, respectively. The BaCoCa (Kück & Struck,

2014) output was manually inspected for the taxon-specific Chi-square test statistic and degrees of freedom for all individuals, as well as the overall taxon-specific RCFV values of each alignment.

The homogeneity test examines amino acid proportions from all species to determine if amino acid content is homogeneous across the genus *Anas* (Kück & Struck, 2014). With the Chi-square homogeneity test the test statistic for each alignment at $\alpha = 0.05$. RCFV values are taxon-specific compositional biases that are evaluated based on the frequency of the amino acid for each species per respective alignment (Kück & Struck, 2014). A high RCFV value indicates that the partition has increased levels of compositional variability (Kück & Struck, 2014). Compositional variable or bias regions are localised segments in gene sequences that are predominantly made up of a variable subset of amino acids across comparative taxa (Harrison, 2006; Zhong *et al.*, 2011; Kück & Struck, 2014). Standard deviation was used to evaluate species-specific RCFV values for each alignment, and alignments with a standard deviation ≥0.005 were selected, yielding 66 alignments (Figure 3.9). Species-specific RCFV fold increase was then examined, with fold increases ≥1× being selected and the function of those SCOs being determined using eggNOG-mapper v.2.0.6 (Huerta-Cepas *et al.,* 2017).

3.3 Results and discussion

3.3.1 Assembly comparative analysis

The genome of the YBD was compared to the genomes of eleven closely related bird taxa. The genome sequences of two wild Mallards, six domesticated Mallards (five Pekin Ducks and one Laying Duck), a single wild and domesticated Mallard hybrid, one wild Eastern Spot-billed Duck and one domestic chicken were obtained from the NCBI (Table 3.1). Seven of the twelve comparative genomes were derived from female birds (with the inclusion of the YBD), while the Pekin-Pooled (GCA_015476345) was derived from a pooled female and male Pekin multi-isolate sample (Table A1; NCBI, 2022). Female Aves are the heterogametic sex (ZW); thus, female sequences represent the entire genome of the respective species (Degrandi *et al.,* 2020a).

The genome size of the *Anas* taxa ranges between 1.049 to 1.210 Gb (Table 3.1), with the YBD genome being the sixth largest. The YBD genome also does not include gapped (Ns) regions, and with the exclusion of these from the other genomes, the YBD genome still ranks as the sixth largest. The genomes of Wild×Dom (GCA_900411745) and Pekin-Female (GCA_000355885), a wild and

domestic cross and a Pekin Duck, respectively, have the smallest genomes, measuring 1.049 and 1.070 Gb, respectively, and have a high gap content, with an average of 575,036.5 Ns per assembly. CAU-Wild (GCA_008746955), a wild Mallard Duck, likewise has an atypical assembly size of 1.12 Gb. The average genome size is comparable to those referenced in the study by Zhang *et al.* (2014a), which looked at 48 bird species across 36 avian orders (including *G. gallus* and the *A. platyrhynchos domesticus* genomes), with a genome size of 1.13 Gb on average.

The Pekin-Female and Wild×Dom assemblies, both built with short reads, have the most and second-most contigs, with 175,076 and 102,646 contigs, respectively. In addition, the two assemblies have the lowest GC content, averaging 40.95%. Using the N50 value to determine contiguity (Table 3.1), the YBD assembly (N50: 351,454) is more contiguous than assemblies sequenced with short reads, with the chicken genome being the most contiguous with the greatest N50 value (22,526,059 bp; Table 3.1). The Wild×Dom and Pekin-Female assemblies, however, are the least contiguous, with N50 values of 35,070 and 34,236 respectively, and were excluded from further study. The YBD assembly also has the longest minimum contig length of 4,074 bp, while comparative genomes including the chicken genome averaged at 290.63 bp.

Table 3.1: Assembly statistics of the Yellow-billed Duck and comparative genomes

Assembly No.	Organism code	Contig No.	Min Contig length (bp)	Max Contig length (bp)	N50	Assembly length (bp)	No. of Ns	GC% [a]
1	*A. undulata*	10767	**4074**	2510228	351454	1161236934	0	42.03
2	*A. zonorhyncha*	71511	31	1346860	117067	1179966058	**680005**	41.89
3	Pekin-Pooled	1516	7	28495473	6168860	1184304312	1461	42.27
4	Pekin-2015	2237	521	12331622	2872511	1090905466	676226	41.53
5	Pekin-Male	41127	48	992079	101284	1080131854	30430	41.42
6	CAU-Pekin	609	511	55383155	16035322	1185139615	155	**42.42**
7	CAU-Laying	793	530	53929205	8766561	1210593445	675	42.55
8	CAU-Wild	1960	207	**73505640**	11896999	1210757631	306	42.41
9	D2L-Wild	62416	26	1777703	146169	1170472851	595622	41.89
10	Wild×Dom	102646	51	363439	35070	1049737960	567600	40.90
11	Pekin-Female	**175076**	51	363439	34236	1070538623	582473	40.99
12	*G. gallus*	660	1214	65391699	**22526059**	1049948559	226	42.20

[a] Obtained from the MOSGA pipeline: GC% from RepeatMasker v.4.1.1 (Smit *et al.,* 2013)

Highest metric for each respective category in bold

Of the twelve compared genomes, seven were sequenced using Pacific Biosciences SMRT long reads (Table A1). BUSCO analysis showed that, on average, the genomes of the comparative genomes was 94.6% complete (Figure 3.1). Relative to this, the BUSCO completeness of our YBD genome (93.6%) is only marginally lower. The lowest was observed for CAU-Wild and hence it was removed from further analyses.

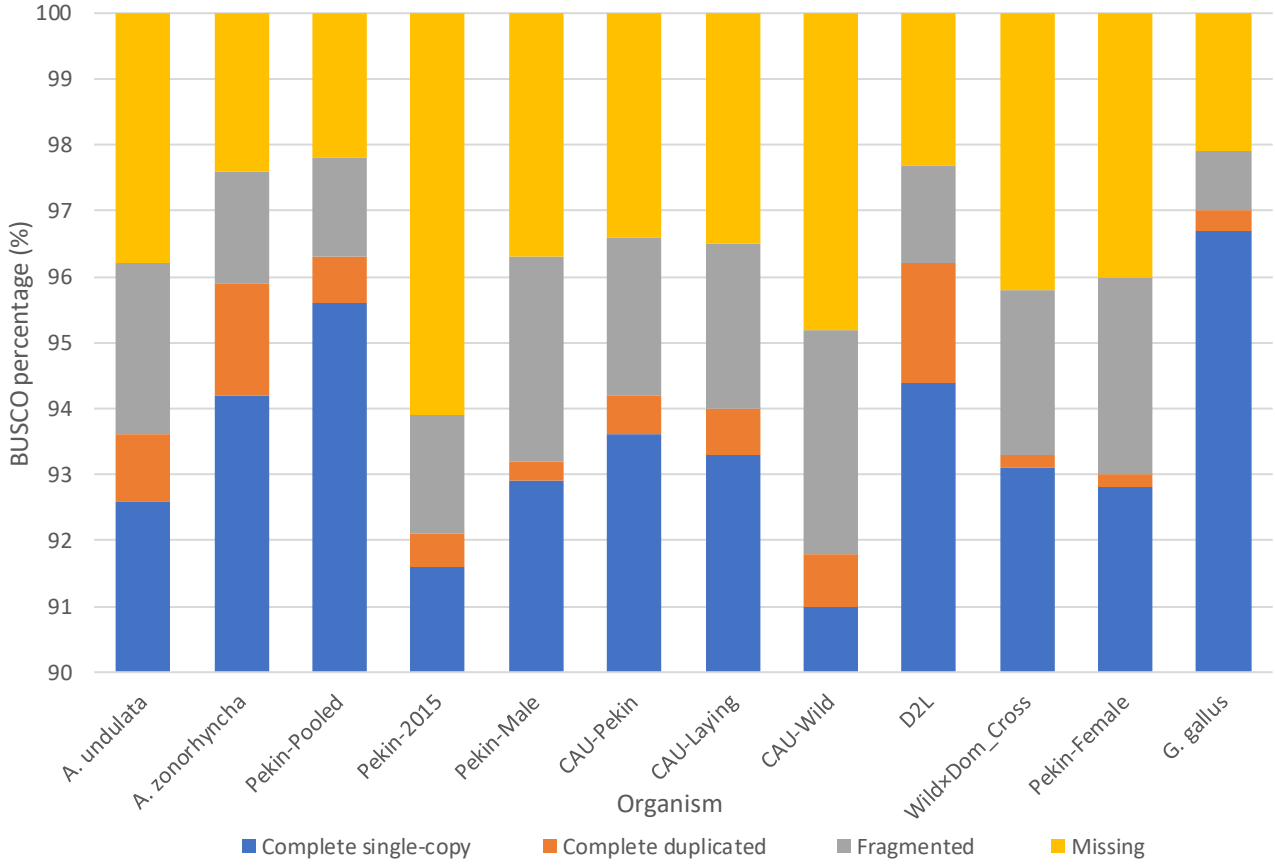


Figure 3.1: Genome assembly completeness assessment using BUSCO v.4.1.4.

3.3.2 *Anas* core genome phylogeny

A total of 7,378 SCOs were found conserved across all eight compared genomes. After alignment of the individual SCOs, concatenation of the alignments and the removal of poorly aligned positions, a final alignment comprising 3,092,402 amino acid positions (290,466 were singleton sites, 2,788,876 constant sites, and 13,060 parsimony informative sites) was used to construct a core genome phylogeny (Figure 3.2). A singleton site is an alignment site that contains no less than two types of amino acids, with at most one having multiple occurrences, whereas a parsimony informative site contains no less than two different sequence character states (Kumar *et al.,* 2016).

The optimal evolutionary model JTT+F+I+G4 (Jones *et al.,* 1992) was used for tree construction (Posada & Crandall, 1998). The JTT model (Jones *et al.,* 1992) is a highly utilised phylogenetic tree inference model. The method computes the differences between sequence pairs from a phylogeny with the lowest phylogenetic distance and a sequence similarity greater than 85% (Jones *et al.,* 1992; Posada & Crandall, 1998; Whelan & Goldman, 2001). The *A. platyrhynchos* ducks largely cluster together, with *A. zonorhyncha* and *A. undulata*. However, one Mallard individual, DL2-Wild clusters with *A. zonorhyncha,* suggesting they are more closely related to each other than to the YBD. This clustering is supported by high bootstrap values.

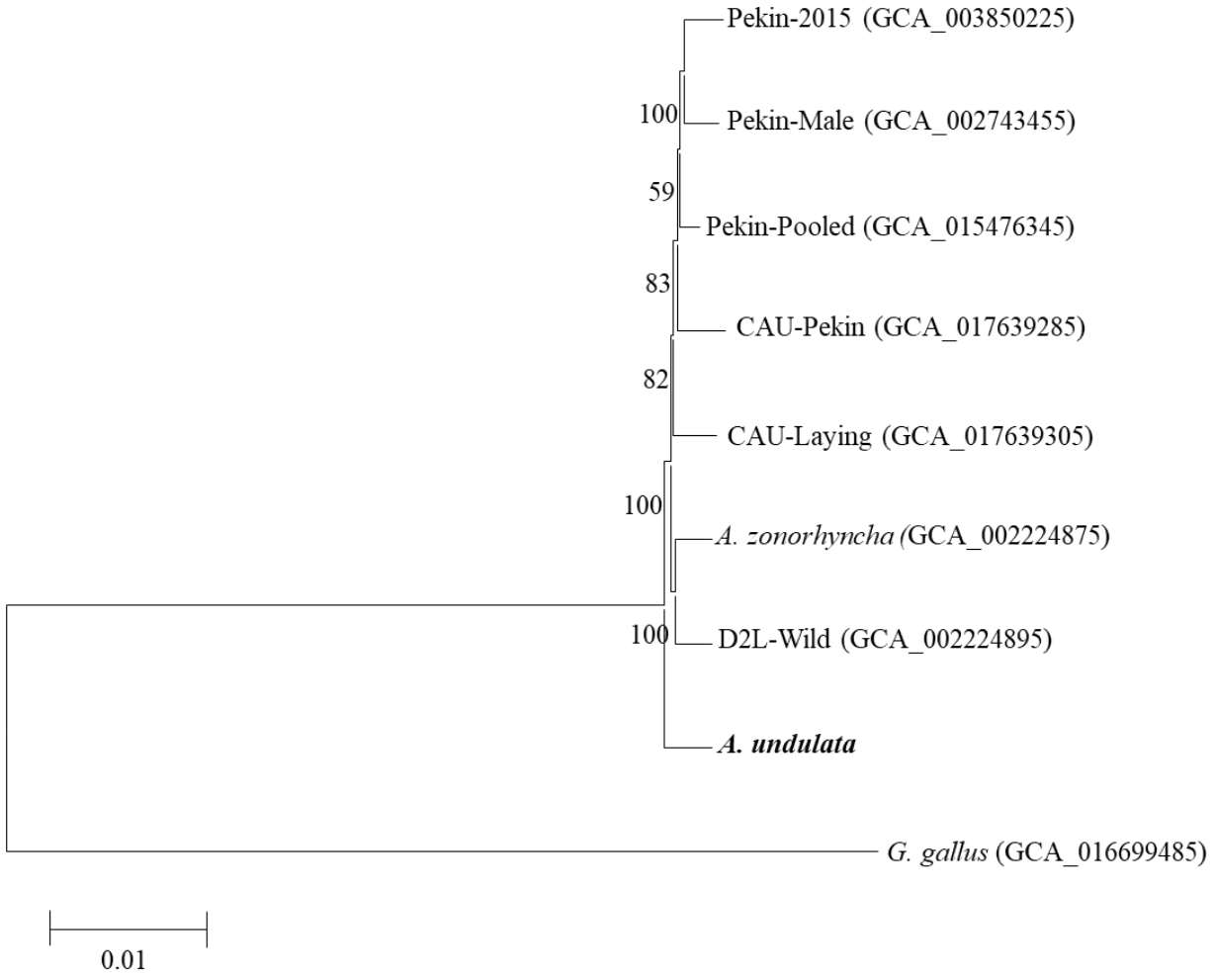


Figure 3.2: Core genome phylogeny of eight ducks in the genus *Anas*. With the exception of D2L-Wild, which clustered with *A. zonorhyncha* and *A. undulata* closer to the root, Mallards were mostly clustered together.

The inclusion of the chicken in the phylogeny resulted in a long branch which largely masked the topology of the members of the genus *Anas.* As such, a core genome phylogeny comprising only the eight comparative ducks was constructed. The curated concatenated alignment comprising 10,526 SCOs consisted of 4,258,157 amino acid positions (86,141 singleton sites, 4,150,703 constant sites and 21,313 parsimony informative sites). The core genome phylogeny was constructed using the optimal evolutionary model LG+I+G (Le & Gascuel, 2008). LG is an expansion of Whelan and Goldman's (2001) WAG matrix, which an approximation of the evolutionary history of a gene or protein sequence considering substitution, transition and transversion probabilities (Le & Gascuel, 2008). A key observation from this phylogeny (Figure 3.3) is that the ducks form two distinct clades, with one clade comprising the wild *Anas* species, while the domesticated Mallard Ducks form a distinct clade. This is despite the fact that the domesticated and wild Mallard Ducks form part of the same species, *A. platyrhynchos*, with Mallard domestication records dating back 2,200 years (Serjeantson, 2014; Zhang *et al.,* 2018; Wang *et al.,* 2020). The clades are further supported by high bootstrap values.

The consideration of the whole genome, as in our phylogeny, suggests the extensive global evolution of genes and proteins post-domestication has resulted in a distinct domesticated duck clade, or species. Zhang *et al*. (2014b) inferred phylogeny using Jarvis *et al*. (2014) methodology, which performed 'total-evidence nucleotide tree' or TENT inference using the ExaML algorithm (Stamatakis & Aberer, 2013). Zhang *et al*. (2014b) is an extension of the Zhang *et al.* (2014a) study, which inferred the phylogenomics of the Zhang *et al.* (2014a) annotated genomes. Zhang *et al*. (2014b) and Jarvis *et al*. (2014), determined that the *A. platyrhynchos domesticus* or Pekin Duck forms the Galloanserae clade with the *G. gallus* and *Meleagris gallopavo* (Turkey).

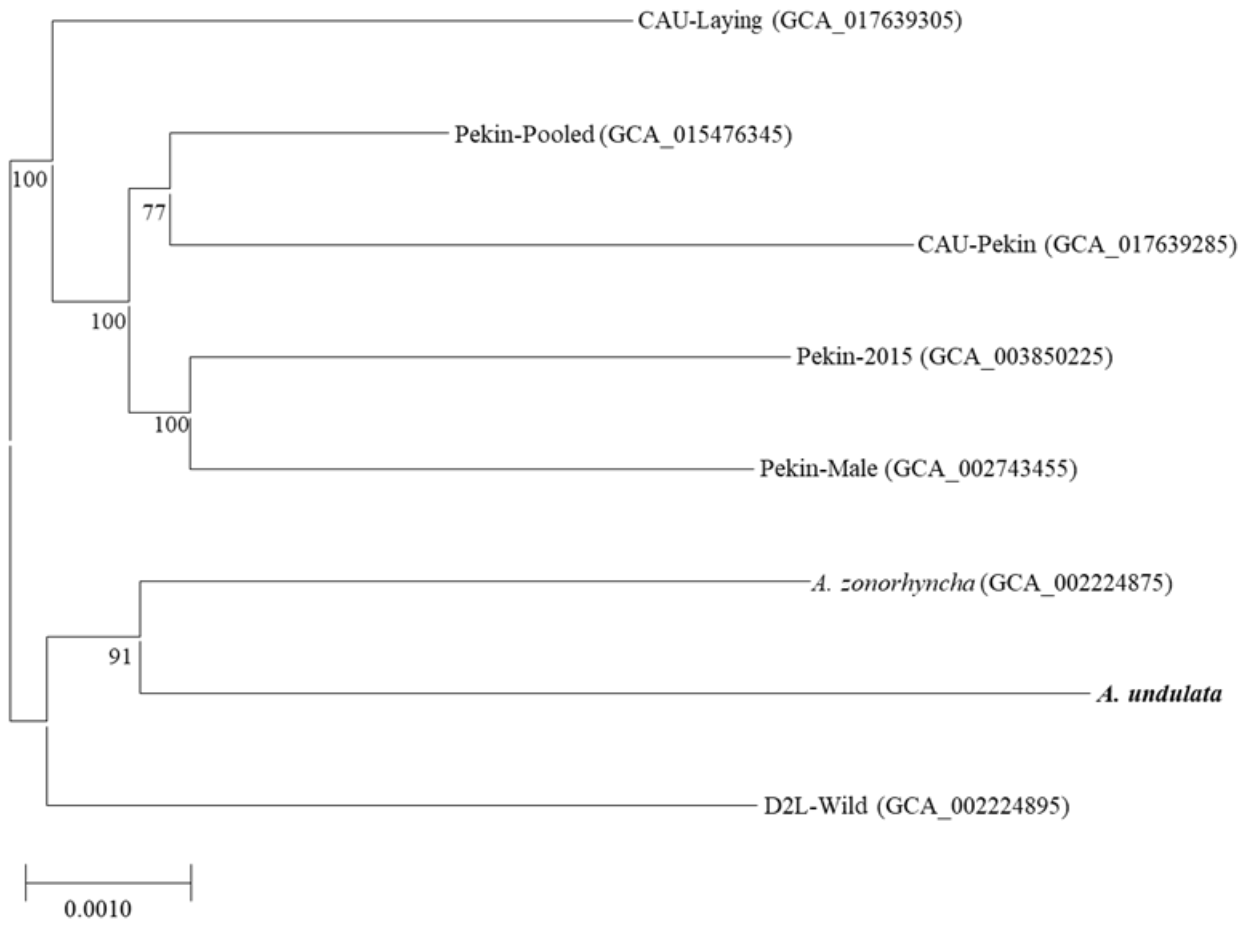


Figure 3.3: Midpoint rooted maximum likelihood phylogeny of the genus *Anas.* All wild *Anas* genomes, including the YBD cluster together, with the YBD and *A. zonorhyncha* being closely related. All domesticated Mallards also cluster together with Pekin Ducks being closely related relative to the Laying Duck.

3.3.3 Annotation comparative analysis

MOSGA identified an average of 35,401 protein coding genes (average gene length = 11321,64 bp) in the genomes of the eight comparative taxa (excluding the chicken) with the YBD containing 39,258 genes. Apart from the Pekin-Male (GCA_002743455), all *Anas* genomes carrying <30,000 genes, with the CAU-Pekin (GCA 017639285) genome contain the most with 39,947 genes. EggNOG-Mapper (Huerta-Cepas *et al*., 2017) effectively annotated 91.70% of the 35,401 average genes, while an alignment against the SwissProt database using DIAMOND assigned functions to 46.51% of the encoded proteins for each genome.

3.3.4 Repeat element comparative analysis

Tandem repeats and transposable element identification and annotation was conducted using RepeatMasker (Smit *et al.,* 2013). The YBD has an above-average repeat composition of 10.17% relative to the *Anas* genome average (at 9.39%). This fits the transposable element content of less than 15% (n = 363) reported by Feng *et al.* (2020) in their dense comparative avian genomic study. Feng *et al.* (2020), however reported outliers, which include *Tricholaema leucomelas* (Acacia Pied Barbet; SAMN12253853), *Psilopogon haemacephalus* (Coppersmith Barbet; SAMN12253782) and the *Eubucco bourcierii* (Red-headed Barbet; SAMN12253768) all with repeats spanning >25% of their genome. A previous study which aimed to evaluate genome content, which included repeat elements across 36 bird orders using 48 birds by Zhang *et al.* (2014a), most of the genomic repeats span 2 to 9%, with the *G. gallus* genome with 9.82% of the genome comprising repeats*, A. platyrhynchos* with 5.85%, and *Picoides pubescens* (Downy Woodpecker) being the outlier composed of 22.15% repeats. The variation coefficient at 14.29% (including the outgroup) indicates medium overall repeat element dispersion (Lopes *et al.,* 2021). With reference to the *Anas* genomes (Figure 3.4), LINEs have the highest genomic composition at an average of 5.33%, The highest proportion of LINEs is observed in the YBD, with 5.68% of the genome comprising this repeat element. LINEs have previously been shown to be the most abundant repeat element, with the CR1 superfamily being the most common in Aves *(*Zhang *et al*. 2014a; Kapusta and Suh, 2017). This is also the case in this study with CR1 having the highest genome composition across all the comparative genomes.

Table 3.2: The gene, functional annotation, tRNA, and rRNA genome summaries for all comparative Aves

Assembly No.	Organism code	Gene number	Mean gene length (bp)	Mean CDS length (bp)	No. of EggNOG annotations	No. of SwissProt annotations	No. of tRNAs	No. of rRNAs
1	*A. undulata*	39258	9840.47	238.71	35842	17230	260	91
2	*A. zonorhyncha*	33789	12524.53	207.54	31651	**17380**	276	47
3	Pekin-Pooled	38651	10795.08	235.81	34776	16432	**277**	51
4	Pekin-2015	33106	11645.48	219.53	30233	14432	268	166
5	Pekin-Male	29721	12415.41	204.02	28181	15839	215	16
6	CAU-Pekin	**39947**	10459.64	**241.54**	**36113**	17253	274	48
7	CAU-Laying	39357	10561.34	238.34	35399	16762	275	**225**
9	D2L-Wild	33983	12331.2	209.6	31702	17148	275	15
12	*G. gallus*	30797	**12527.9**	198.76	28253	15707	211	118

Highest metric for each respective category in bold

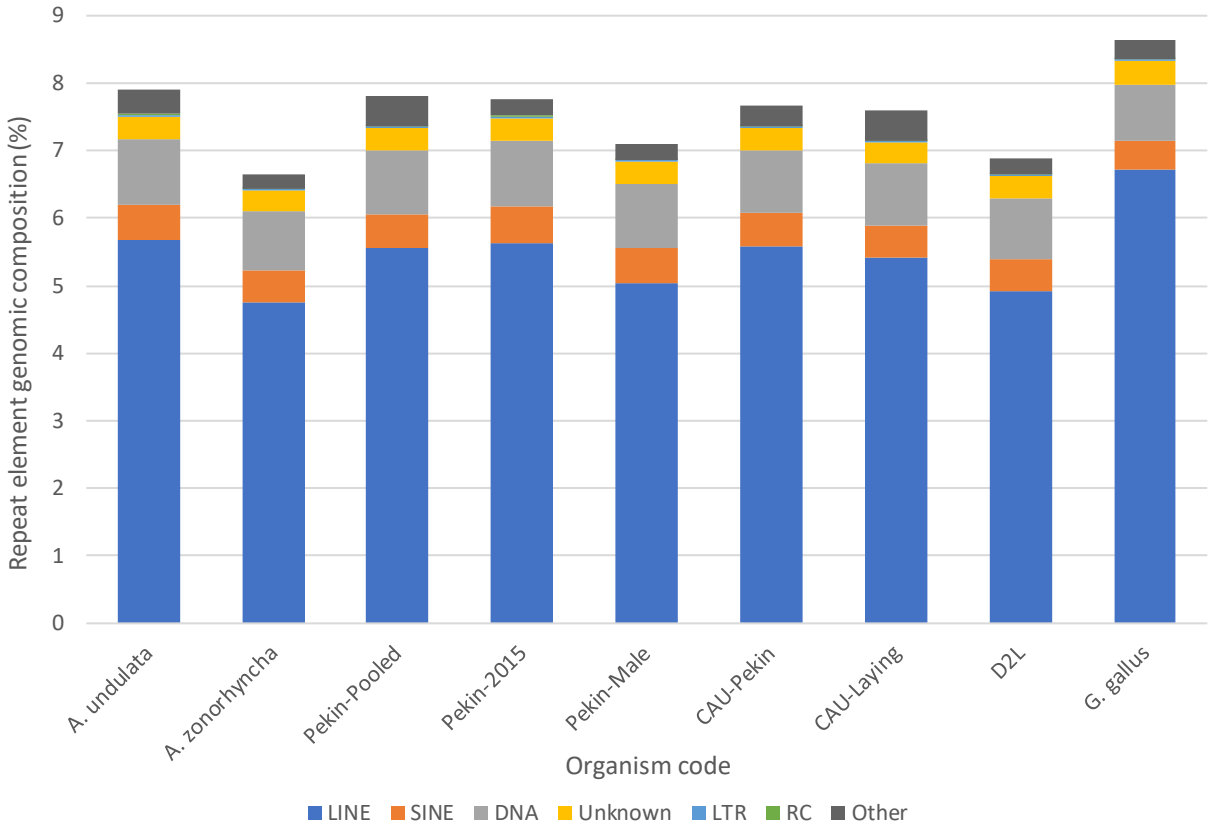


Figure 3.4: Repeat element genomic composition across nine respective genomes. The other repeat category includes simple repeats, low complexity regions, satellites and small RNAs.

3.3.5 Pangenome analysis

The OrthoFinder output was used to determine the core, accessory, and unique elements of the pan-genome of the genus *Anas* (Figure 3.5). The core genome comprises 11,239 orthogroups which are shared by all taxa. A total of 3,943 orthogroups are accessory to the genus *Anas* (i.e. not in the chicken genome), while 737 are unique to *G. gallus*. The *A. zonorhyncha* and *A. platrhynchos* have the highest number of accessory orthogroups at 668, this followed by the YBD and *A. zonorhyncha* with 196 shared orthogroups. On average, the genus *Anas* contains 252 unique orthogroups, with the highest orthogroup content belonging to the YBD. When evaluating individual Mallards, the D2L-Wild genome has the highest number of unique orthorgroup (117), followed by Pekin-Male (90), Pekin-2015 (72), CAU-Laying (50), Pekin-Pooled (46) and lastly CAU-Pekin (43).



Figure 3.5: The genus *Anas* pangenome, with the *G. gallus* as the outgroup. All Mallard Duck genomes were concatenated with values representing the shared SCOs.

3.3.6 Paralogue analysis

Orthologues and paralogues are two distinct genes that evolved independently, with orthologues developing from a single ancestral gene and paralogues emerging from gene duplication (Koonin, 2005). The difference between the two homologs is essential for the evolutionary categorization of genes as well as the reliable functional annotation of novel genomes (Koonin, 2005). Paralogous genes, appear multiple times within a genome caused by a duplication event, which can subsequently result in gain-of-function (Koonin, 2005; Qian & Zhang, 2008; Saalbach, 2022). The primary source of novel gene function development is duplication occurrences, with gene retention within a particular population assumed to be owing to genetic drift and positive selection (Qian & Zhang, 2008). Positive selection to raise gene dosage, which is the proportional gene copy number within a genome, has been postulated to drive gene retention (Qian & Zhang, 2008; Basilicata & Keller Valsecchi, 2021). The gene dosage effect hypothesis, which will be deduced through genus

Anas paralogue analysis, describes phenotypic alterations caused by increased gene copy number (Basilicata & Keller Valsecchi, 2021).

With an average paralogue copy number of 51.04 per protein, YBD has the highest average copy number, indicating high duplication rates compared to D2L-Wild at 29.66 and *A. zonorhyncha* at 26.5 (Figure 3.6). Most notably, the YBD genome had the highest gene copy number in 35 of the top 50 orthogroups, seen in Figure 3.6, which generated the large standard deviation. The genes encoding olfactory receptors, have the greatest standard deviation, with the YBD having a 3.53-fold increase in gene copy average compared to the other wild *Anas* birds. This could be attributed to short-read sequencing used to sequence the *A. platyrynchus* D2L-Wild. genomes, as Driver and Balakrishnan (2021) discovered gene copy number underreporting of the olfactory receptor gene family in short-read assemblies, whereas long-read assemblies had high resolution, allowing for increased gene copy identification. Figure 3.7, which maps YBD gene copy number against domestic ducks, demonstrates the same trend, with one of the long-read assemblies, that of *A. platyrynchus* CAU-Pekin, having the highest copy number in this orthogroup (630 copies) and a short-read assembly, fix as above Pekin-Male, having the lowest copy number (123 copies).

The autophagy-related Atg4 protein, a cysteine protease, is present in 105 copies in the YBD genome, none in the D2L-Wild genome, and only one in the *A. zonorhyncha* genome. Atg4 is the only Atg8-specific protease enzyme, which catalyses the delipidation of Atg8-phosphatidylethanolamine of the complex between the amine group of the phospholipid and the C-terminal carboxyl moiety of Atg8 for it to be active in autophagy (Li *et al.,* 2017; Maruyama & Noda, 2018). Autophagy is a conserved and regulated intracellular mass degradation pathway within eukaryotes that is significantly triggered by environmental stressors like starvation or infections (Li *et al.,* 2017; Barna *et al.,* 2018; Maruyama & Noda, 2018). According to the gene dosage hypothesis, greater Atg4 gene copies indicate that the YBD has enhanced Atg8-phosphatidylethanolamine delipidation, where Atg8 can be recycled, hence upregulating the stimulation of membrane elongation during autophagosome maturation (Li *et al.,* 2017; Maruyama & Noda, 2018). Upregulated autophagosome maturation suggests that the YBD has been exposed to more environmental stressors, resulting in the positive selection of increased Atg4 copy numbers.

The enhanced heat shock transcription factor copy number in comparison to the two other wild *Anas* genomes further supports autophagy upregulation and increased environmental tolerance for the YBD. The copy number of the transcription factor is 12× greater for the YBD than the D2L-Wild and *A. zonorhyncha* copy number average of 2.5. The constitutive transcription factor, which increases the expression of genes encoding heat shock proteins that repair heat-injured intracellular proteins (Barna *et al.,* 2018), has 30 copies in the YBD, two in the D2L-Wild, and three in the *A. zonorhyncha*. High copy numbers of the constitutive transcription factor suggest high heat tolerance, and it has been proven that autophagy is activated in numerous mammalian cell lines and genetic models following heat stress (Barna *et al.,* 2018). This reaffirms that the YBD may have been subjected to higher environmental stresses, and that positive selection has resulted in increased copy number, resulting in increased environmental resilience.

Next the top 50 orthogroups with the highest standard deviation when comparing the YBD and domestic *Anas* genomes were evaluated (Figure 3.7). The average copy number of genes varies across domestic Aves, with the CAU-Pekin genome containing the greatest average at 73.68, the Pekin-Male genome having the lowest at 30.02, and the YBD having a median average of 60.92. The YBD has the highest gene copy number in 12 of the 50 orthogroups, with the OG0000214 orthogroup having the highest fold increase (13×), with 40 copies compared to the domestic copy number average of 3 copies. The orthogroup had no functional hit, however an SPOC domain, follows a similar trend with 40 YBD gene copies and the domestic copy number averaging at 4,4 copies. The SPOC domain, also known as the Spen orthologue and paralogue C-terminal domain, is a eukaryotic conserved fifteen to twenty kDa protein domain that forms a deformed β-barrel from several α-helices and seven β-strands (Sánchez-Pulido *et al.,* 2004; Appel *et al.,* 2023). SPOC-containing proteins are frequently linked to cellular differentiation and development, as well as transcriptional regulation (Sánchez-Pulido *et al.,* 2004; Appel *et al.,* 2023). The SPOC domain of plant homeodomain finger protein 3 (PHF3) is required for the functional execution of transcriptional and mRNA stabilisation in neural genes, whereas a PHF3 paralogue, Death inducer obliterator (DIDO), is implicated in gene splicing and stem cell development (Appel *et al.,* 2023). This suggests that the YBD may have enhanced and stabilised neuronal gene transcription, as well as increased stem cell reserves, which necessitate more SPOC domain copies to allow transcription protein binding.

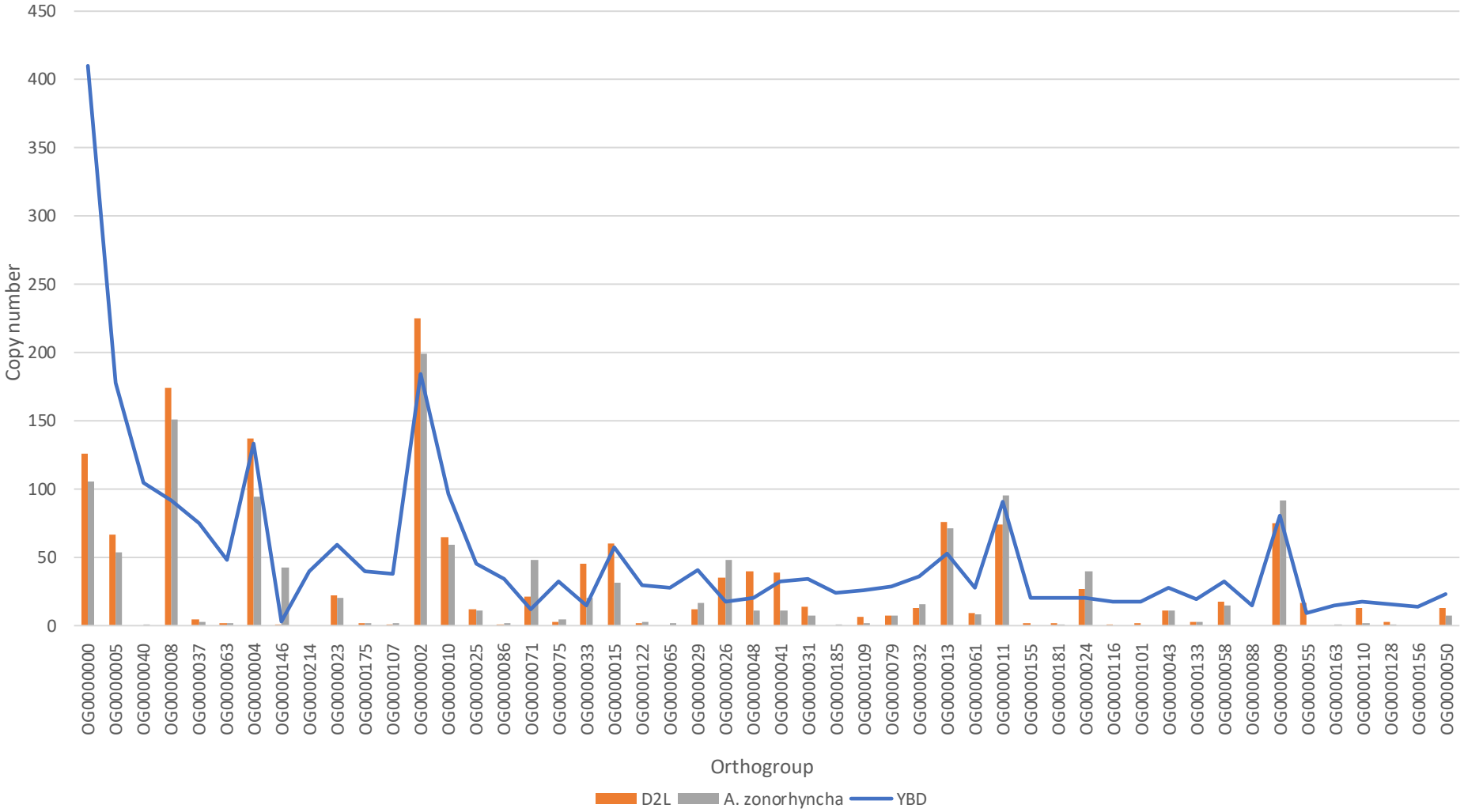


Figure 3.6: Paralogous genes of wild *Anas* genomes. The top 50 orthogroups with the biggest standard deviation when comparing the YBD paralogue copy number to the other wild *Anas* genomes.

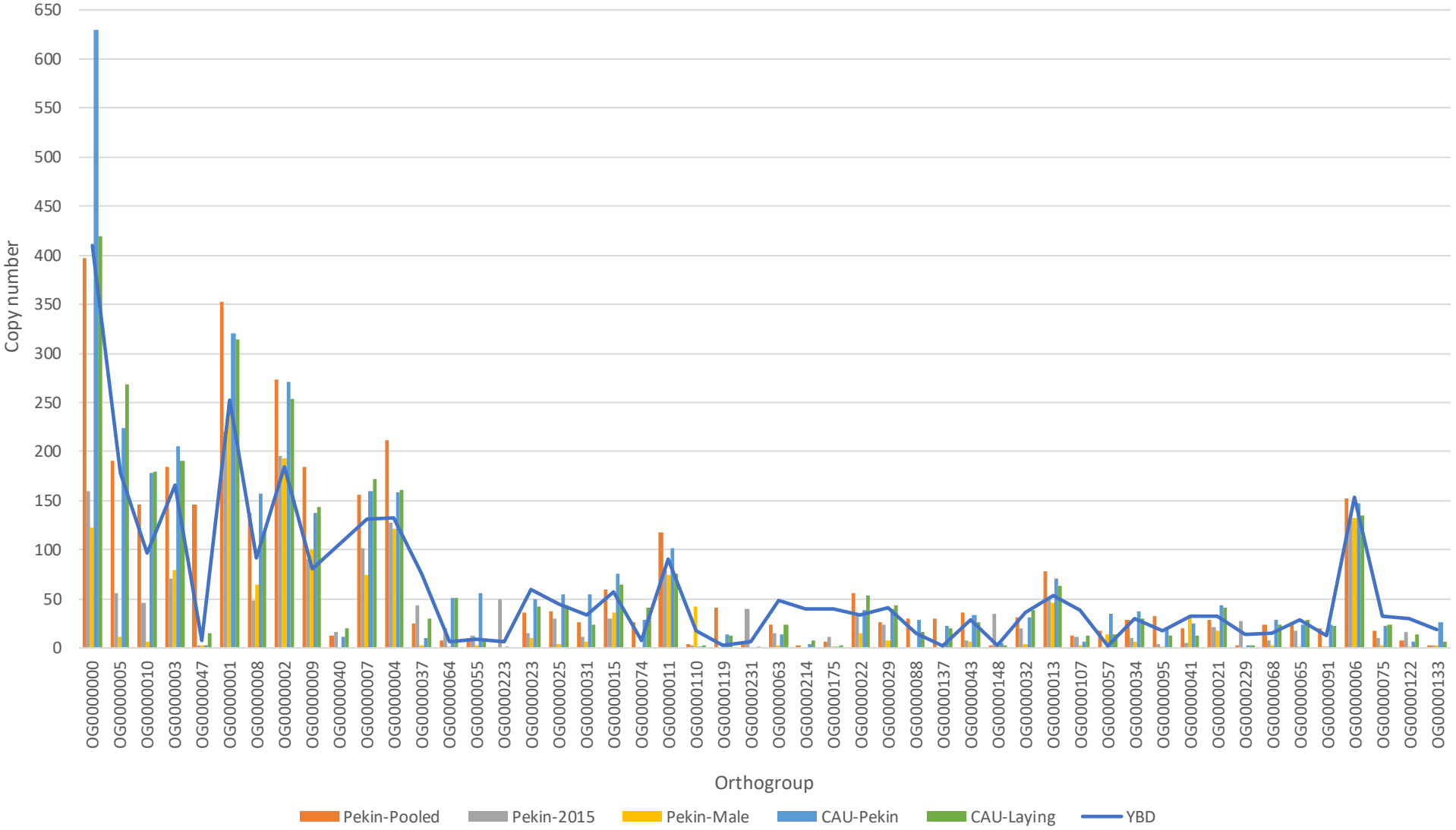


Figure 3.7: Paralogous comparative analysis of the YBD and domestic *Anas* genomes.

Genes coding for an immunoglobulin with a V-type domain had an 8.1× higher copy number in the domesticated ducks compared to the YBD. Within the orthogroup, the YBD has two copies, whereas domestic birds averaged 16.2 copies, with CAU-Pekin having the largest copy number with 35 copies and Pekin-2015 having the lowest with no copies. Since copy number is not preserved between domestic birds, gene copy variations within this functional grouping can be attributed to assembly differences. Next, wild and domesticated gene copy variation was examined to acquire a better understanding of how domestication affects gene copy counts (Figure 3.8). Despite, no copy number conservation across all the comparative taxa, the domestic ducks contain a higher copy number average in 33 of the 50 orthogroups (Figure 3.8). Higher gene copy numbers in domestic ducks indicate higher duplication rates, which resulted in higher gene numbers. This then modified the phenotype of the domesticated duck, and this phenomenon, as previously said, underpins evolution and adaptation (Koonin, 2005; Qian & Zhang, 2008; Saalbach, 2022). To investigate protein diversification further as a result of evolutionary adaptation, codon bias was examined.

3.3.7 Codon diversity identification using BaCoCa

BaCoCa determined that at 133 degrees of freedom, the Chi-square test statistic for each alignment averaged 2.28, with the greatest statistic value being 87.47. The tool found that, given the default significance level of 0.01 for the respective comparative SCOs they are not significantly deviating from homogeneity. This was expected as data stems from the same genus, thus reflecting that the core genome phylogeny of the genus *Anas* is homogenous. The overall RCFV value of each species across an alignment was used to examine compositional bias that may have been introduced during tree construction. For each alignment the standard deviation of the RCFV values was assessed and alignments with a standard deviation ≥0.005 were selected. This resulted in SCO 66 alignments depicted in Figure 3.9, where burgundy are high RCFV values. To successfully extract heatmap hotspots, RCFV values with a ≥1 fold increase per alignment relative to other comparative taxa were selected.

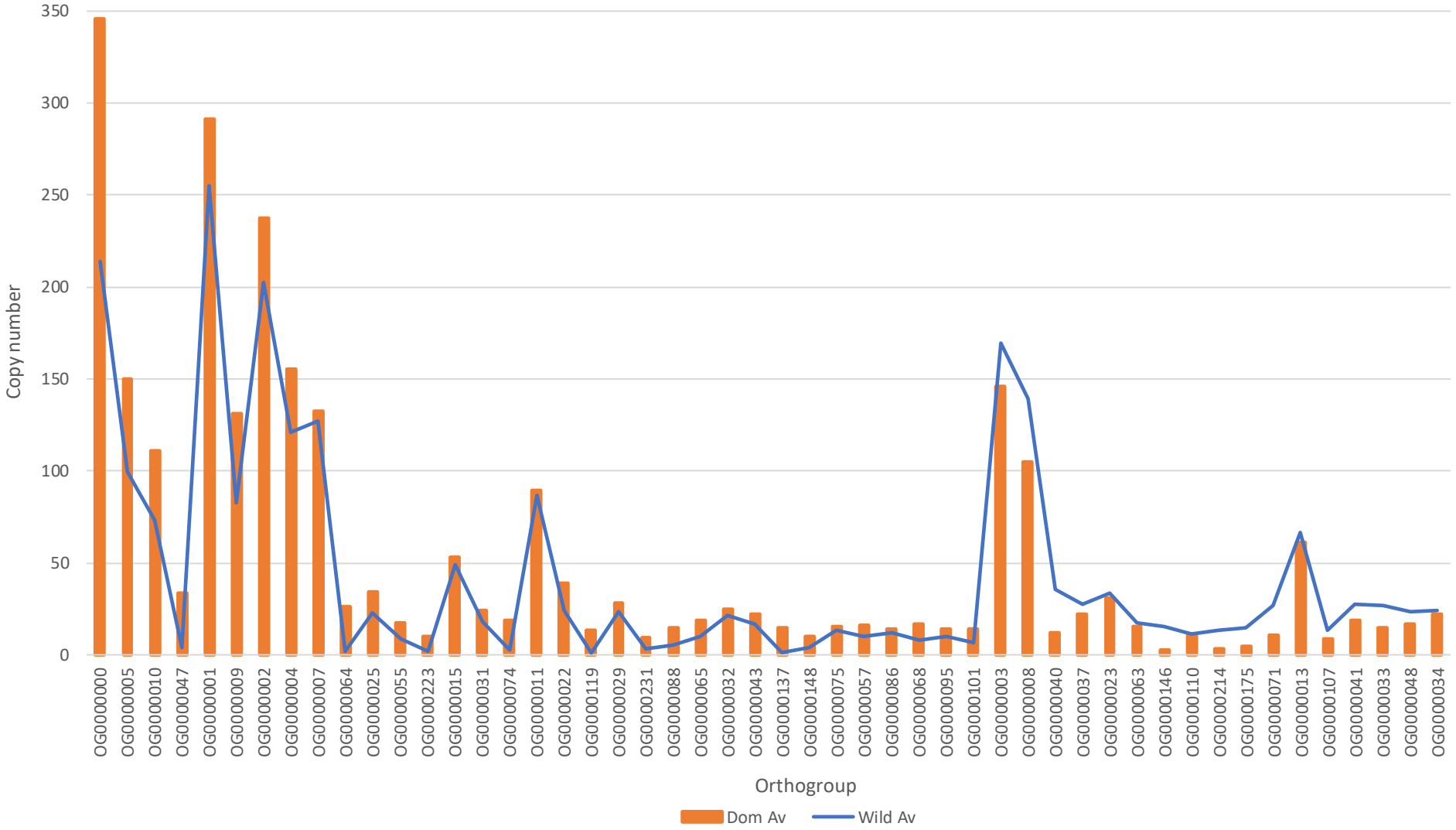


Figure 3.8: Paralogous comparative analysis of wild and domestic *Anas* genomes

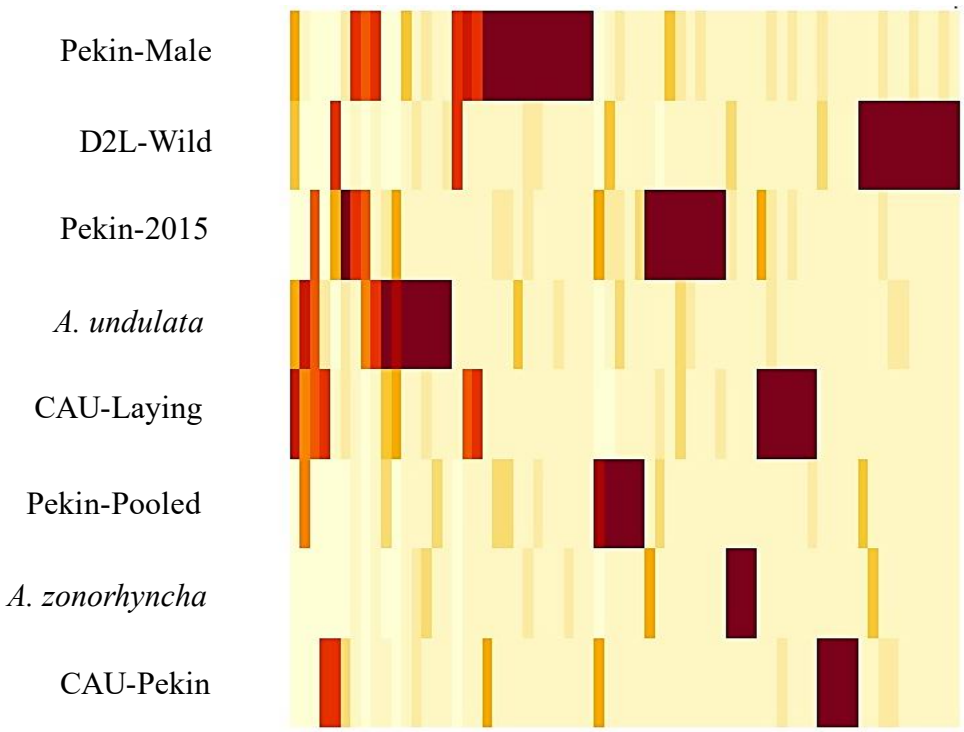


Figure 3.9: Species-specific RCFV heatmap. Data was obtained from 66 SCO alignments with a standard deviation ≥0.005, with burgundy representing high RCFV values.

A total of twenty proteins had RCFV values above the threshold for the YBD compared to the other ducks, with fold increases ranging from 1.27 to 6.96; fifteen of which were functionally annotated using eggNOG-mapper v.2.0.6 (Huerta-Cepas *et al.,* 2017). YBD genes with distinct amino acid compositions have functions ranging from Ribonuclease H protein (1.27-fold increase) to the X1 isoform of the NF-kappa-B inhibitor-interacting Ras-like protein 2 or NKIRAS2 (6.96-fold increase). The Ribonuclease H is a widely conserved eukaryotic enzyme involved in DNA replication and repair (Cerritelli & Crouch, 2009). The X1 NKIRAS2 isoform suppresses the activation of NF-kappa-B, an M1 macrophage transcription factor implicated in the initiation of an inflammatory response (Sarais *et al.,* 2020). Except for the Pekin-Pooled duck, the function of all these genes is comparable the comparative birds despite the significant RCFV fold increase, which indicates amino acid diversity. This pattern is repeated in other birds, demonstrating that despite amino acid and codon diversity, the function of the genes is conserved, and therefore despite compositional variability, there is minimal functional diversity. These species-specific localised amino acid and codon-diverse regions should thus serve as a starting point for selection analysis.

3.4 Conclusion

The YBD genome assembly metrics and MOSGA annotations are comparable to high confidence *Anas* genomes. Not only that, but the genomic data of this study is comparable to earlier avian genetic investigations. The first *Anas* phylogenomic analysis revealed the clade division of domesticated and wild ducks, which can serve as the starting point to further investigate *Anas* species domestication. This can be used as a starting point to investigate domestication within the genus. Despite having the smallest genome among the wild ducks, the YBD has the greatest average gene copy number, which includes high copy numbers for autophagy-related genes implicated in environmental adaptation and resistance. This suggests that the YBD may be more resistant to environmental changes as a result of high gene duplication for resistance genes. Domesticated ducks, on the other hand, exhibited a greater average copy number, implying that high duplication rates may be linked to domestication. An investigation into amino acid and codon diversification on the shared genes discovered twenty genes where the YBD has high levels of diversity in comparison to the other comparative birds; this pattern is consistent in other birds, and this phenomenon can serve as a steppingstone for selection analysis. Future research should look not only at *Anas* domestication, but also at what promotes the phenomenon within the clade and how different genetic regions influence wild and domesticated duck diversity.

Chapter 4: Summary and concluding remarks

4.1 Project summary

Aves have the highest species diversity within the amniotes clade, making them a focus of evolutionary, speciation, and adaption research (Koonin, 2005; Qian & Zhang, 2008; Ksepka, 2021 Saalbach, 2022). Of particular importance the Yellow-billed Duck (YBD), a nocturnal omnivorous dabbling duck from the genus *Anas* of the Anseriformes order (Livezey, 1991). Dispersive nomads native to 22 eastern and southern African countries, with a high South African localisation, with the grassland and fynbos biomes having the highest observed population (Oatley & Prys-Jones, 1986; Maclean, 1997Brown *et al.,* 2019; de Souza *et al.,* 2019; BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). The duck population provides ecosystem services, which are organism-based activities indirectly or directly beneficial to humans (Sekercioglu, 2006; Suri *et al.,* 2017). This includes avitourism's economic contribution, wetland nutrient deposition and aquatic endo and ectozoochoric propagule dispersal (Sekercioglu, 2006; Damschen *et al.,* 2008; Holyoak *et al.,* 2008; Department of Trade and Industry, 2010; Reynolds, 2016; Suri *et al.,* 2017). Threats to the bird include recreational hunting, avian botulism and hybridisation which threaten genetic integrity, assessed using non-YBD specific markers (Brown *et al.,* 2019; de Souza *et al.,* 2019; Stephens *et al.,* 2019; BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). With the above-mentioned high South African localisation and the potential to generate YBD-specific markers for conservation efforts, the aim of this study was to sequence, assemble, and annotate the first Yellow-billed Duck genome, followed by comparative genomic analysis with other species in the genus *Anas*.

To accomplish this, a female blood sample from the de Souza *et al.* (2019) study was ethically collected from Barberspan Nature Reserve, for phenol/chloroform gDNA extraction. Following PacBio sequencing to generate CLR and HiFi reads, *de novo* whole genome assembly via multiple rounds of Canu long read genome assembly and Racon polishing was conducted to yield 1.16 Gb assembly with 93.6% completeness. MOSGA, or Modular Open-Source Genome Annotator, identified 39,258 genes and functionally annotated over 90% (36,046) of them, while also identifying 91 rRNA and 424 tRNA genes and that repeats span 10.17% of the genome. In the third chapter, eight NCBI-obtained *Anas* genomes were evaluated based on genome content, phylogeny post orthology inference, pangenome analysis, paralogy, and codon diversity analysis, with the fully MOSGA annotated genomes serving as the baseline. When evaluating *Anas*

phylogeny, one of the significant results in Chapter 3 is the separation of domesticated and wild ducks. The Aves also have paralogy and amino acid diversification across the genus *Anas*. As a result of the individual genes with diverse amino acid composition, sequencing more genomes to increase population size and provide a more comprehensive understanding of the distinctions between the YBD, Mallard, and *A. zonorhyncha*.

The generation of more avian genomes is expected to result in a shift toward posteriori marker selection, in which loci are retrieved from genome-wide alignments, allowing for more sensitive marker selection, a better understanding of avian phylogenetics, and the genetic processes that enable functional and phenotypic diversity (Bravo *et al.,* 2021). This study's Yellow-billed Duck sequence can aid in the development of the first YBD-specific microsatellite markers to facilitate hybridisation monitoring. Previous studies like the Brown *et al*. (2019), de Souza *et al*. (2019), and Stephens *et al.* (2019) studies have used established non-YBD specific microsatellites to study the hybridisation phenomenon, hence the creation of novel, Yellow-billed Duck microsatellites is crucial. Microsatellites identified post gene sequencing, thus, to establish new markers whole genome sequencing is required to conduct YBD-specific marker development in the monitoring and subsequent conservation of all these dabbling ducks. The avian genome can also be employed in research other than conservation, such as virology and toxicity (Huang *et al.,* 2013; Flores-Santin & Burggren, 2021), adaptation (Schaefer & Böhning-Gaese, 2008; Cheng *et al.,* 2021), and colourisation (Campagna *et al.,* 2017; Knief *et al.,* 2019; Baiz *et al.,* 2021).

This study can also be the extension of a pigmentation gene diversity study that investigates bill colouration and plumage diversity among genus *Anas* species. The Yellow-billed Duck, most Pekin Ducks, and Eastern Spot-billed Duck possess variations of a yellow bill, with the two former ducks having the entire bill yellow and the latter only having the tip of the bill yellow (BirdLife International, 2022; Southern African Bird Atlas Project 2, 2022). The yellow bills are most likely due to the presence of carotenoids, which substantially absorb light in the blue-green spectrum, resulting in either yellow, orange, or red phenotypic traits (Walsh *et al.,* 2012; Stavenga *et al.*, 2017). Carotenoids must be received through diet because Aves cannot synthesize the chemical (Stavenga *et al.,* 2017). Because all three birds are omnivorous and have comparable diets, it is obvious that carotenoid-based coloration is likewise dependent on the underlying genetic pathways that control carotenoid metabolism (Walsh *et al.,* 2012; Stavenga *et al.,* 2017). Thus, in

combination with the YBD genome, selective carotenoid accumulation through pigment metabolism of the genus *Anas* can be evaluated, as well as how pigment deposition is controlled by different cells resulting in localised colouration. Further studies should focus on the variation that exists between the YBD and the Eastern Spot-billed duck and the similarities that exist between the YBD and most Pekin Ducks.

References

1. Abdurakhmonov, I.Y., 2016. Introduction to microsatellites: basics, trends and highlights. Microsatellite markers, 1, p.13.
2. Agreement on the Conservation of African-Eurasian Migratory Waterbirds (AEWA), 2018. MOP7: Agreement Text and Annexes. 7th Session of the Meeting of the Parties to AEWA. Durban, South Africa: UNEP/AEWA Secretariat, pp.21-36.
3. Alhakami, H., Mirebrahim, H. and Lonardi, S., 2017. A comparative evaluation of genome assembly reconciliation tools. Genome biology, 18(1), pp.1-14.
4. Allendorf, F.W., Leary, R.F., Spruell, P. and Wenburg, J.K., 2001. The problems with hybrids: setting conservation guidelines. Trends in Ecology and Evolution, 16(11), pp.613-622.
5. Al-Okaily, A.A., 2016. HGA: de novo genome assembly method for bacterial genomes using high coverage short sequencing reads. BMC Genomics, 17(1), pp.1-11.
6. Anderson, E., 1948. Hybridization of the habitat. Evolution, pp.1-9.
7. Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.
8. Appel, L.M., Franke, V., Benedum, J., Grishkovskaya, I., Strobl, X., Polyansky, A., Ammann, G., Platzer, S., Neudolt, A., Wunder, A. and Walch, L., 2023. The SPOC domain is a phosphoserine binding module that bridges transcription machinery with co- and post-transcriptional regulators. Nature Communications, 14(1), p.166.
9. Authority, Department of Environmental Affairs and Tourism of the Republic of South Africa, 2003. Agreement on the Conservation of African-Eurasian Migratory Waterbirds (The Hague, 1995): First report by South Africa on the implementation of the Agreement.
10. Bairoch, A., Boeckmann, B., Ferro, S. and Gasteiger, E., 2004. Swiss-Prot: juggling between evolution and stability. Briefings in Bioinformatics, 5(1), pp.39-55.
11. Baiz, M.D., Wood, A.W., Brelsford, A., Lovette, I.J. and Toews, D.P., 2021. Pigmentation genes show evidence of repeated divergence and multiple bouts of introgression in Setophaga warblers. Current Biology, 31(3), pp.643-649.
12. Bajorath, J., Hinrichs, W. and Saenger, W., 1988. The enzymatic activity of proteinase K is controlled by calcium. European Journal of Biochemistry, 176(2), pp.441-447.
13. Barna, J., Csermely, P. and Vellai, T., 2018. Roles of heat shock factor 1 beyond the heat shock response. Cellular and Molecular Life Sciences, 75, pp.2897-2916.

14. Basilicata, M.F. and Keller Valsecchi, C.I., 2021. The good, the bad, and the ugly: Evolutionary and pathological aspects of gene dosage alterations. PLoS Genetics, 17(12), p.e1009906.
15. Benjamini, Y. and Speed, T.P., 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research, 40(10), pp.e72-e72.
16. Bennett, J. and Whitten, S., 2003. Duck hunting and wetland conservation: Compromise or synergy?. Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie, 51(2), pp.161-173.
17. Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W., 2012. GenBank. Nucleic Acids Research, 41(D1), pp.D36-D42.
18. Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M. and Phillippy, A.M., 2015. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nature Biotechnology, 33(6), pp.623-630.
19. BirdLife International (2022). IUCN Red List for birds. [online] Available at: http://www.birdlife.org [Accessed: 04 June 2022].
20. Biscotti, M.A., Olmo, E. and Heslop-Harrison, J.S., 2015. Repetitive DNA in eukaryotic genomes. Chromosome Research, 23, pp.415-420.
21. Bravo, G.A., Schmitt, C.J. and Edwards, S.V., 2021. What have we learned from the first 500 avian genomes. Annu Rev Ecol Evol Syst, 52(1), pp.611-639.
22. Brown, J.I., Lavretsky, P., Cumming, G.S. and Peters, J.L., 2019. Strong population structure and limited gene flow between Yellow-billed Ducks and Mallards in southern Africa. The Condor, 121(4), p.duz042.
23. Brown, R.B. and Audet, J., 2008. Current techniques for single-cell lysis. Journal of the Royal Society Interface, 5(suppl_2), pp.S131-S138.
24. Brüniche-Olsen, A., Kellner, K.F., Belant, J.L. and DeWoody, J.A., 2021. Life-history traits and habitat availability shape genomic diversity in birds: implications for conservation. Proceedings of the Royal Society B, 288(1961), p.20211441.
25. Buchfink, B., Xie, C. and Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods, 12(1), pp.59-60.

26. Campagna, L., Repenning, M., Silveira, L.F., Fontana, C.S., Tubaro, P.L. and Lovette, I.J., 2017. Repeated divergent selection on pigmentation genes in a rapid finch radiation. Science Advances, 3(5), p.e1602404.
27. Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J., 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Molecular Biology and Evolution, 38(12), pp.5825-5829.
28. Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution, 17(4), pp.540-552.
29. Cerritelli, S.M. and Crouch, R.J., 2009. Ribonuclease H: the enzymes in eukaryotes. The FEBS journal, 276(6), pp.1494-1505.
30. Chan, P.P. and Lowe, T.M., 2016. GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes. Nucleic Acids Research, 44(D1), pp.D184-D189.
31. Chen, Y.C., Liu, T., Yu, C.H., Chiang, T.Y. and Hwang, C.C., 2013. Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PloS one, 8(4), p.e62856.
32. Cheng, H.H., Warren, W.C. and Zhou, H., 2021. Avian genomics. In: C. G. Scanes and S. Dridi, eds. Sturkie's Avian Physiology. 7th ed. Oxford: Elsevier, pp. 7-16.
33. Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E. and Turner, S.W., 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nature Methods, 10(6), pp.563-569.
34. Chin, C.S., Peluso, P., Sedlazeck, F.J., Nattestad, M., Concepcion, G.T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A. and Cramer, G.R., 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods, 13(12), pp.1050-1054.
35. Chu, J., Mohamadi, H., Warren, R.L., Yang, C. and Birol, I., 2017. Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. Bioinformatics, 33(8), pp.1261-1270.

36. Chum, O., Philbin, J. and Zisserman, A., 2008. Near duplicate image detection: Min-hash and TF-IDF weighting. In Bmvc, 810, pp. 812-815.
37. Conrad, K.F., Robertson, R.J. and Boag, P.T., 2000. Difficulties storing and preserving tyrant flycatcher blood samples used for genetic analyses. The Condor, 102(1), pp.191-193.
38. Cortes, C., Mohri, M., Riley, M. and Rostamizadeh, A., 2008. Sample selection bias correction theory. In Algorithmic Learning Theory: 19th International Conference, ALT 2008, Budapest, Hungary, October 13-16, 2008. Proceedings 19 (pp. 38-53). Springer Berlin Heidelberg.
39. Cuesta, S.M., Rahman, S.A., Furnham, N. and Thornton, J.M., 2015. The classification and evolution of enzyme function. Biophysical Journal, 109(6), pp.1082-1086.
40. Damschen, E.I., Brudvig, L.A., Haddad, N.M., Levey, D.J., Orrock, J.L. and Tewksbury, J.J., 2008. The movement ecology and dynamics of plant communities in fragmented landscapes. Proceedings of the National Academy of Sciences, 105(49), pp.19078-19083.
41. Dawson, D.A., Ball, A.D., Spurgin, L.G., Martín-Gálvez, D., Stewart, I.R., Horsburgh, G.J., Potter, J., Molina-Morales, M., Bicknell, A.W., Preston, S.A. and Ekblom, R., 2013. High-utility conserved avian microsatellite markers enable parentage and population studies across a wide range of species. BMC Genomics, 14(1), pp.1-22.
42. de Souza, S.G., Symes, C.T., Smit-Robinson, H. and Mollett, J.M., 2019. Minimal evidence of interspecific hybridisation between the Yellow-billed Duck and introduced Mallard in central and northwestern South Africa. Ostrich, 90(4), pp.285-301.
43. Degrandi, T.M., Barcellos, S.A., Costa, A.L., Garnero, A.D., Hass, I. and Gunski, R.J., 2020. Introducing the bird chromosome database: An overview of cytogenetic studies in birds. Cytogenetic and Genome Research, 160(4), pp.199-205.
44. Degrandi, T.M., Gunski, R.J., Garnero, A.D.V., Oliveira, E.H.C.D., Kretschmer, R., Souza, M.S.D., Barcellos, S.A. and Hass, I., 2020b. The distribution of 45S rDNA sites in bird chromosomes suggests multiple evolutionary histories. Genetics and Molecular Biology, 43.
45. Department of Trade and Industry (the DTI) of the Republic of South Africa, 2010. Avitourism in South Africa. Research and Analysis Report. Pretoria

46. Donà, F. and Houseley, J., 2014. Unexpected DNA loss mediated by the DNA binding activity of ribonuclease A. PLoS One, 9(12), p.e115008.
47. Donald, P.F., Fishpool, L.D., Ajagbe, A., Bennun, L.A., Bunting, G., Burfield, I.J., Butchart, S.H., Capellan, S., Crosby, M.J., Dias, M.P. and Diaz, D., 2019. Important Bird and Biodiversity Areas (IBAs): The development and characteristics of a global inventory of key sites for biodiversity. Bird Conservation International, 29(2), pp.177-198.
48. Driver, R.J. and Balakrishnan, C.N., 2021. Highly contiguous genomes improve the understanding of avian olfactory receptor repertoires. Integrative and Comparative Biology, 61(4), pp.1281-1290.
49. Dyomin, A.G., Koshel, E.I., Kiselev, A.M., Saifitdinova, A.F., Galkina, S.A., Fukagawa, T., Kostareva, A.A. and Gaginskaya, E.R., 2016. Chicken rRNA gene cluster structure. PloS One, 11(6), p.e0157464.
50. Eddy, S.R. and Durbin, R., 1994. RNA sequence analysis using covariance models. Nucleic Acids Research, 22(11), pp.2079-2088.
51. Eddy, S.R., 1992. HMMER user's guide. Department of Genetics, Washington University School of Medicine, 2(1), p.13.
52. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B. and Bibillo, A., 2009. Real-time DNA sequencing from single polymerase molecules. Science, 323(5910), pp.133-138.
53. Emms, D.M. and Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome biology, 20, pp.1-14.
54. Eo, S.H., Doyle, J.M. and DeWoody, J.A., 2011. Genetic diversity in birds is associated with body mass and habitat type. Journal of Zoology, 283(3), pp.220-226.
55. Ewing, B. and Green, P., 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Research, 8(3), pp.186-194.
56. Fan, W., Xu, L., Cheng, H., Li, M., Liu, H., Jiang, Y., Guo, Y., Zhou, Z. and Hou, S., 2018. Characterization of duck (*Anas platyrhynchos*) short tandem repeat variation by population-scale genome resequencing. Frontiers in Genetics, 9, p.520.

57. Fasulo, D., Halpern, A., Dew, I. and Mobarry, C., 2002. Efficiently detecting polymorphisms during the fragment assembly process. Bioinformatics, 18(suppl_1), pp.S294-S302.
58. Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q.I., Reeve, A.H., Xie, D., Chen, G., Guo, C., Faircloth, B.C. and Petersen, B., 2020. Dense sampling of bird diversity increases power of comparative genomics. Nature, 587(7833), pp.252-257.
59. Flores-Santin, J. and Burggren, W.W., 2021. Beyond the chicken: alternative avian models for developmental physiological research. Frontiers in Physiology, p.1699.
60. Funk, E.R. and Taylor, S.A., 2019. High-throughput sequencing is revealing genetic associations with avian plumage color. The Auk, 136(4), p.ukz048.
61. Furuno, M., Kasukawa, T., Saito, R., Adachi, J., Suzuki, H., Baldarelli, R., Hayashizaki, Y. and Okazaki, Y., 2003. CDS annotation in full-length cDNA sequence. Genome Research, 13(6b), pp.1478-1487.
62. G, Garrett, R. H. & Grisham, C. M., 2016. Biochemistry. In: G, R. H. Garrett & C. M. Grisham, eds. Chapter 29: Transcription and the Regulation of Gene Expression. 6th ed. Boston: Cengage Learning, pp. 1038-1076.
63. Galperin, M.Y. and Koonin, E.V., 2010. From complete genome sequence to 'complete' understanding?. Trends in biotechnology, 28(8), pp.398-406.
64. Galperin, M.Y., Makarova, K.S., Wolf, Y.I. and Koonin, E.V., 2015. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Research, 43(D1), pp.D261-D269.
65. Geldenhuys, G., Hoffman, L.C. and Muller, N., 2013. Gamebirds: A sustainable food source in Southern Africa?. Food Security, 5(2), pp.235-249.
66. GenBank Overview [Internet], n.d. National Center for Biotechnology Information. National Library of Medicine (U.S). Available at: https://www.ncbi.nlm.nih.gov/genbank/ [Accessed: July 28, 2022].
67. Gene Ontology, C., 2015. Gene ontology consortium: going forward. vol. 43, pp.D1049-56. Nucleic Acids Research.
68. Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004. Available at: https://www.ncbi.nlm.nih.gov/genome/ [Accessed 27 August 2022].

69. Genome 10K Community of Scientists, 2009. Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. Journal of Heredity, 100(6), pp.659-674.
70. Genovart, M., 2009. Natural hybridization and conservation. Biodiversity and Conservation, 18(6), pp.1435-1439.
71. Giani, A.M., Gallo, G.R., Gianfranceschi, L. and Formenti, G., 2020. Long walk to genomics: History and current approaches to genome sequencing and assembly. Computational and Structural Biotechnology Journal, 18, pp.9-19.
72. Gill, F., Donsker, D. and Rasmussen, P. eds., 2021. IOC world bird list (v12. 2). IOC.
73. Go, A.C. and Civetta, A., 2020. Hybrid incompatibilities and transgressive gene expression between two closely related subspecies of Drosophila. Frontiers in Genetics, 11, p.599292.
74. Goffeau, A., Barrell, B.G., Bussey, H., Davis, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., Johnston, M. and Louis, E.J., 1996. Life with 6000 genes. Science, 274(5287), pp.546-567.
75. Gogarten, J.P. and Olendzenski, L., 1999. Orthologs, paralogs and genome comparisons. Current opinion in genetics & development, 9(6), pp.630-636.
76. Govaert, L., De Meester, L., Spaak, P. and Hairston Jr, N.G., 2021. Eco-Evolutionary Dynamics in Freshwater Systems.
77. Govender, D., Russell, I.A., Randall, R.M. and Zimmerman, D., 2019. Outbreak of avian botulism and its effect on waterbirds in the Wilderness Lakes, South Africa. Koedoe: African Protected Area Conservation and Science, 61(1), pp.1-13.
78. Government of the Republic of South Africa, 2014. National Environmental Management: Biodiversity Act (10/2004): Alien and Invasive Species Regulations
79. Grant, B.R. and Grant, P.R., 1996. High survival of Darwin's finch hybrids: effects of beak morphology and diets. Ecology, 77(2), pp.500-509.
80. Grant, P.R. and Grant, B.R., 1992. Hybridization of bird species. Science, 256(5054), pp.193-197.
81. Grant, P.R. and Grant, B.R., 1997. Hybridization, sexual imprinting, and mate choice. The American Naturalist, 149(1), pp.1-28.

82. Grant, P.R. and Grant, B.R., 2010. Conspecific versus heterospecific gene exchange between populations of Darwin's finches. Philosophical Transactions of the Royal Society B: Biological Sciences, 365(1543), pp.1065-1076.
83. Grant, P.R. and Grant, B.R., 2019. Hybridization increases population variation during adaptive radiation. Proceedings of the National Academy of Sciences, 116(46), pp.23216-23224.
84. Grant, P.R., Grant, B.R., Keller, L.F., Markert, J.A. and Petren, K., 2003. Inbreeding and interbreeding in Darwin's finches. Evolution, 57(12), pp.2911-2916.
85. Greco, M., Saez, C.A., Brown, M.T. and Bitonti, M.B., 2014. A simple and effective method for high quality co-extraction of genomic DNA and total RNA from low biomass Ectocarpus siliculosus, the model brown alga. PloS One, 9(5), p.e96470.
86. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A. and Eddy, S.R., 2003. Rfam: an RNA family database. Nucleic Acids Research, 31(1), pp.439-441.
87. Gu, S. and Jin, R., 2012. Assembly and function of the botulinum neurotoxin progenitor complex. In: A. Rummel and T. Binz, eds., Botulinum Neurotoxins. Heidelberg: Springer, pp.21-44.
88. Harrison, P.M., 2006. Exhaustive assignment of compositional bias reveals universally prevalent biased regions: analysis of functional associations in human and Drosophila. BMC bioinformatics, 7, pp.1-15.
89. Harrison, R.G. and Larson, E.L., 2014. Hybridization, introgression, and the nature of species boundaries. Journal of Heredity, 105(S1), pp.795-809.
90. Hill, K. and Smith, T., 2012. Genetic diversity within Clostridium botulinum serotypes, botulinum neurotoxin gene clusters and toxin subtypes. In: A. Rummel and T. Binz, eds., Botulinum Neurotoxins. Heidelberg: Springer, pp.1-20.
91. Hoff, K.J. and Stanke, M., 2019. Predicting genes in single genomes with AUGUSTUS. Current Protocols in Bioinformatics, 65(1), p.e57.
92. Holyoak, M., Casagrandi, R., Nathan, R., Revilla, E. and Spiegel, O., 2008. Trends and missing parts in the study of movement ecology. Proceedings of the National Academy of Sciences, 105(49), pp.19060-19065.

93. Hon, T., Mars, K., Young, G., Tsai, Y.C., Karalius, J.W., Landolin, J.M., Maurer, N., Kudrna, D., Hardigan, M.A., Steiner, C.C. and Knapp, S.J., 2020. Highly accurate long-read HiFi sequencing data for five complex genomes. Scientific Data, 7(1), pp.1-11.
94. Hoysak, D.J. and Ankney, D.C., 1996. Correlates of behavioral dominance in mallards and American black ducks. Animal Behaviour, 51(2), pp.409-419.
95. Hu, T., Chen, G., Xu, Z., Luo, S., Wang, H., Li, C., Shan, L. and Zhang, B., 2022. De novo Whole-Genome Sequencing and Assembly of the Yellow-Throated Bunting (Emberiza elegans) Provides Insights into Its Evolutionary Adaptation. Animals, 12(15), p.2004.
96. Huang, Y., Li, Y., Burt, D.W., Chen, H., Zhang, Y., Qian, W., Kim, H., Gan, S., Zhao, Y., Li, J. and Yi, K., 2013. The duck genome and transcriptome provide insight into an avian influenza virus reservoir species. Nature Genetics, 45(7), pp.776-783.
97. Hubbs, C.L., 1955. Hybridization between fish species in nature. Systematic Zoology, 4(1), pp.1-20.
98. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F. and Wheeler, T.J., 2016. The Dfam database of repetitive DNA families. Nucleic Acids Research, 44(D1), pp.D81-D89.
99. Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C. and Bork, P., 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular Biology and Evolution, 34(8), pp.2115-2122.
100. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J. and von Mering, C., 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research, 47(D1), pp.D309-D314.
101. IUCN, 2022. The IUCN Red List of Threatened Species. Version 2022-2. https://www.iucnredlist.org. [Accessed: March 03, 2022].
102. Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y., Faircloth, B.C., Nabholz, B., Howard, J.T. and Suh, A., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science, 346(6215), pp.1320-1331.
103. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B., 2007. Genome-wide mapping of in vivo protein-DNA interactions. Science, 316(5830), pp.1497-1502.

104. Johnson, E.A. and Bradshaw, M., 2001. *Clostridium botulinum* and its neurotoxins: a metabolic and cellular perspective. Toxicon, 39(11), pp.1703-1722.
105. Jones, D.T., Taylor, W.R. and Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Bioinformatics, 8(3), pp.275-282.
106. Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research, 42(D1), pp.D199-D205.
107. Kanz, C. et al., 2005. The EMBL Nucleotide Sequence Database. Nucleic Acids Research, 33, D29–D33
108. Kapusta, A. and Suh, A., 2017. Evolution of bird genomes—a transposon's-eye view. Annals of the New York Academy of Sciences, 1389(1), pp.164-185.
109. Katinas, L., Morrone, J.J. and Crisci, J.V., 1999. Track analysis reveals the composite nature of the Andean biota. Australian Journal of Botany, 47(1), pp.111-130.
110. Kear, J. ed., 2005. Ducks, geese and swans: general chapters, species accounts (Anhima to Salvadorina) (Vol. 1). Oxford University Press.
111. Khosravinia, H., Murthy, H.N., Parasad, D.T. and Pirany, N., 2007. Optimizing factors influencing DNA extraction from fresh whole avian blood. African Journal of Biotechnology, 6(4).
112. Kliman, R.M., 2016. Encyclopedia of evolutionary biology. Academic Press.
113. Knief, U., Bossu, C.M., Saino, N., Hansson, B., Poelstra, J., Vijay, N., Weissensteiner, M. and Wolf, J.B., 2019. Epistatic mutations under divergent selection govern phenotypic variation in the crow hybrid zone. Nature Ecology and Evolution, 3(4), pp.570-576.
114. Koonin, E.V., 2005. Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet., 39, pp.309-338.
115. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research, 27(5), pp.722-736.
116. Kriventseva, E.V., Rahman, N., Espinosa, O. and Zdobnov, E.M., 2007. OrthoDB: the hierarchical catalog of eukaryotic orthologs. Nucleic Acids Research, 36(suppl_1), pp.D271-D275.

117. Kronenberg, Z., 2020. Direct phased genome assembly using Nighthawk on HiFi reads. [Blog] PacBio Blog, Available at: https://www.pacb.com/blog/direct-phased-genome-assembly-using-nighthawk-on-hifi-reads/ [Accessed: 3 August 2022].
118. Ksepka, D., 2021. Evolution of birds. In: C. Scanes and S. Dridi, eds. Sturkie's Avian Physiology, 7th ed. Oxford: Elsevier, pp.83-108
119. Kück, P. and Struck, T.H., 2014. BaCoCa–A heuristic software tool for the parallel assessment of sequence biases in hundreds of gene and taxon partitions. Molecular phylogenetics and evolution, 70, pp.94-98.
120. Kulmuni, J. and Westram, A.M., 2017. Intrinsic incompatibilities evolving as a by-product of divergent ecological selection: Considering them in empirical studies on divergence with gene flow.
121. Lamichhaney, S., Berglund, J., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C.J., Wang, C., Zamani, N. and Grant, B.R., 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature, 518(7539), pp.371-375.
122. Lang, D., Zhang, S., Ren, P., Liang, F., Sun, Z., Meng, G., Tan, Y., Li, X., Lai, Q., Han, L. and Wang, D., 2020. Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. Gigascience, 9(12), p.giaa123.
123. Lawson, D.M., Williams, C.K., Lavretsky, P., Howell, D.L. and Fuller, J.C., 2021. Mallard–Black Duck Hybridization and Population Genetic Structure in North Carolina. The Journal of Wildlife Management, 85(6), pp.1243-1255.
124. Le, S.Q. and Gascuel, O., 2008. An improved general amino acid replacement matrix. Molecular biology and evolution, 25(7), pp.1307-1320.
125. Li, J., Zhang, J., Liu, J., Zhou, Y., Cai, C., Xu, L., Dai, X., Feng, S., Guo, C., Rao, J. and Wei, K., 2021. A new duck genome reveals conserved and convergently evolved chromosome architectures of birds and mammals. Gigascience, 10(1), p.giaa142.
126. Li, M., Fu, Y., Yang, Z. and Yin, X.M., 2017. Measurement of the activity of the Atg4 cysteine proteases. Methods in enzymology, 587, pp.207-225.
127. Li, M., Sun, C., Xu, N., Bian, P., Tian, X., Wang, X., Wang, Y., Jia, X., Heller, R., Wang, M. and Wang, F., 2022. *De novo* assembly of 20 chicken genomes reveals the

undetectable phenomenon for thousands of core genes on microchromosomes and subtelomeric regions. Molecular biology and evolution, 39(4), p.msac066.

128. Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K. and Li, S., 2010. *De novo* assembly of human genomes with massively parallel short read sequencing. Genome Research, 20(2), pp.265-272.

129. Liao, X., Li, M., Zou, Y., Wu, F.X. and Wang, J., 2019. Current challenges and solutions of *de novo* assembly. Quantitative Biology, 7(2), pp.90-109.

130. Lin, Y., Yuan, J., Kolmogorov, M., Shen, M.W., Chaisson, M. and Pevzner, P.A., 2016. Assembly of long error-prone reads using de Bruijn graphs. Proceedings of the National Academy of Sciences, 113(52), pp.E8396-E8405.

131. Lister, R., O'Malley, R.C., Tonti-Filippini, J., Gregory, B.D., Berry, C.C., Millar, A.H. and Ecker, J.R., 2008. Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell, 133(3), pp.523-536.

132. Little, R.M., Vester, K.C. and Crowe, T.M., 1995. Temporal and spatial patterns of breeding activity of 12 duck species (*Anatidae*) in the Cape Provinces, South Africa, and their implications for hunting seasons. South African Journal of Wildlife Research, 25(1), pp.17-22.

133. Liu, J., Wang, Z., Li, J., Xu, L., Liu, J., Feng, S., Guo, C., Chen, S., Ren, Z., Rao, J. and Wei, K., 2021. A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes. Genome Research, 31(3), pp.497-511.

134. Liu, S. and Hansen, M.M., 2017. PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. Molecular Ecology Resources, 17(4), pp.631-641.

135. Livezey, B.C., 1991. A phylogenetic analysis and classification of recent dabbling ducks (Tribe *Anatini*) based on comparative morphology. The Auk, 108(3), pp.471-507.

136. Logsdon, G.A., Vollger, M.R. and Eichler, E.E., 2020. Long-read human genome sequencing and its applications. Nature Reviews Genetics, 21(10), pp.597-614.

137. Maak, S., Wimmers, K., Weigend, S. and Neumann, K., 2003. Isolation and characterization of 18 microsatellites in the Peking duck (*Anas platyrhynchos*) and their application in other waterfowl species. Molecular Ecology Notes, 3(2), pp.224-227.

138. Maclean, G., 1993. Roberts' Birds of Southern Africa. 6th ed. Cape Town: Trustees of the John Voelcker Bird Book Fund, pp.126-127.
139. Maclean, G.L., 1997. Yellowbilled duck. The Atlas of southern African birds, 1, pp.126-127.
140. Madge, S. and Burn, H., 1988. Wildfowl: an identification guide to the ducks, geese and swans of the world. A&C Black.
141. Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. and Ravikesavan, R., 2013. Gene duplication as a major force in evolution. Journal of genetics, 92(1), pp.155-161.
142. Mallet, J., 2005. Hybridization as an invasion of the genome. Trends in Ecology and Evolution, 20(5), pp.229-237.
143. Martin, R., Dreßler, H., Hattab, G., Hackl, T., Fischer, M.G. and Heider, D., 2021a. MOSGA 2: Comparative genomics and validation tools. Computational and Structural Biotechnology Journal, 19, pp.5504-5509.
144. Martin, R., Hackl, T., Hattab, G., Fischer, M.G. and Heider, D., 2021b. MOSGA: modular open-source genome annotator. Bioinformatics, 36(22-23), pp.5514-5515.
145. Maruyama, T. and Noda, N.N., 2018. Autophagy-regulating protease Atg4: structure, function, regulation and inhibition. The Journal of antibiotics, 71(1), pp.72-78.
146. Mashima, J., Kodama, Y., Fujisawa, T., Katayama, T., Okuda, Y., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y. and Takagi, T., 2016. DNA Data Bank of Japan. Nucleic Acids Research, p.gkw1001.
147. Mathew, C.G.P., 1884. The Isolation of high molecular weight eukaryotic DNA. In: J. Walker, ed., Methods in Molecular Biology: Nucleic Acids, 2nd vol. Clifton, New Jersey: Humana press, pp.31-34.
148. McCarthy, E.M., 2006. Handbook of avian hybrids of the world. Oxford university press.
149. McCracken, K.G. and Wilson, R.E., 2011. Gene flow and hybridization between numerically imbalanced populations of two duck species in the Falkland Islands. PLoS One, 6(8), p.e23173.
150. McDonald, A.G. and Tipton, K.F., 2021. Enzyme nomenclature and classification: The state of the art. The FEBS Journal.
151. McGirr, J.A. and Martin, C.H., 2020. Ecological divergence in sympatry causes gene misexpression in hybrids. Molecular Ecology, 29(14), pp.2707-2721.

152. Mendelsohn, S.L. and Young, D.A., 1978. Inhibition of ribonuclease Efficacy of sodium dodecyl sulfate, diethyl pyrocarbonate, proteinase K and heparin using a sensitive ribonuclease assay. Biochimica et Biophysica Acta (BBA)-Nucleic Acids and Protein Synthesis, 519(2), pp.461-473.
153. Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D. and Gurevich, A., 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics, 34(13), pp.i142-i150.
154. Miller, D.B., 1977. Social displays of Mallard Ducks (*Anas platyrhynchos*): effects of domestication. Journal of Comparative and Physiological Psychology, 91(2), p.221.
155. Miller, J.R., Delcher, A.L., Koren, S., Venter, E., Walenz, B.P., Brownley, A., Johnson, J., Li, K., Mobarry, C. and Sutton, G., 2008. Aggressive assembly of pyrosequencing reads with mates. Bioinformatics, 24(24), pp.2818-2824.
156. Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A. and Lanfear, R., 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Molecular biology and evolution, 37(5), pp.1530-1534.
157. Mölder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., Köster, J., 2021. Sustainable data analysis with Snakemake. F1000Res 10, 33.
158. Moran, B.M., Payne, C., Langdon, Q., Powell, D.L., Brandvain, Y. and Schumer, M., 2021. The genomic consequences of hybridization. ELife, 10, p.e69016.
159. Morneau, D., 2021.Pan-genomes: moving beyond the reference. Nature Milestones: Genomic Sequencing, 15, p. S19.
160. Muller, W., 2008. Hybridisation, and the conservation of the grey duck in New Zealand.
161. Myers, G., 2014. Efficient local alignment discovery amongst noisy long reads. In International Workshop on Algorithms in Bioinformatics, pp. 52-67.
162. Nadachowska-Brzyska, K., Burri, R., Smeds, L. and Ellegren, H., 2016. PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white Ficedula flycatchers. Molecular ecology, 25(5), pp.1058-1072.

163. Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M. and Snyder, M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 320(5881), pp.1344-1349.

164. Nawrocki, E.P. and Eddy, S.R., 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 29(22), pp.2933-2935.

165. Ncbi.nlm.nih.gov. n.d. VecScreen: Screen for Vector Contamination. [online] Available at: https://www.ncbi.nlm.nih.gov/tools/vecscreen/ [Accessed: 25 June 2022].

166. Ng, S.B., Turner, E.H., Robertson, P.D., Flygare, S.D., Bigham, A.W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E.E. and Bamshad, M., 2009. Targeted capture and massively parallel sequencing of 12 human exomes. Nature, 461(7261), pp.272-276.

167. Ngcaba, N, Department of Environmental Affairs of the Republic of South Africa, 2015. Agreement on the Conservation of African-Eurasian Migratory Waterbirds (The Hague, 1995): Report on the implementation of AEWA for the period 2012-2014.

168. Nikaein, H. and Sharifi-Zarchi, A., 2022. Alignment of Single-Molecule Sequencing Reads by Enhancing the Accuracy and Efficiency of Locality-Sensitive Hashing. bioRxiv.

169. Notredame, C., Higgins, D.G. and Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology, 302(1), pp.205-217.

170. Nussinov, R., Tsai, C.J. and Jang, H., 2019. Protein ensembles link genotype to phenotype. PLoS computational biology, 15(6), p.e1006648.

171. Oatley, T.O. and Prys-Jones, R., 1986. A comparative analysis of movements of southern African waterfowl (*Anatidae*), based on ringing recoveries. South African Journal of Wildlife Research, 16(1), pp.1-6.

172. Ohtsubo, Y., Sakai, K., Nagata, Y. and Tsuda, M., 2019. Properties and efficient scrap-and-build repairing of mechanically sheared 3'DNA ends. Communications Biology, 2(1), pp.1-9.

173. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D. and Astashyn, A., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), pp.D733-D745.

174. Ottenburghs, J., 2019. Multispecies hybridization in birds. Avian Research, 10(1), pp.1-11.
175. Ottenburghs, J., Ydenberg, R.C., van Hooft, W.F., Van Wieren, S.E. and Prins, H.H., 2015. The Avian Hybrids Project: gathering the scientific literature on avian hybridization. Ibis, 157(4), pp.892-894.
176. Paulke, E. and Haase, E., 1978. A comparison of seasonal changes in the concentrations of androgens in the peripheral blood of wild and domestic ducks. General and Comparative Endocrinology, 34(4), pp.381-390.
177. Petren, K., 1998. Microsatellite primers from Geospiza fortis and cross-species amplification in Darwin's finches. Molecular Ecology, 7(12), pp.1782-1784.
178. Petren, K., Grant, B.R. and Grant, P.R., 1999. Low extrapair paternity in the cactus finch
179. Pirazzini, M., Rossetto, O., Eleopra, R. and Montecucco, C., 2017. Botulinum neurotoxins: biology, pharmacology, and toxicology. Pharmacological reviews, 69(2), pp.200-235.
180. Posada, D. and Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics (Oxford, England), 14(9), pp.817-818.
181. Price-Waldman, R. and Stoddard, M.C., 2021. Avian coloration genetics: recent advances and emerging questions. Journal of Heredity, 112(5), pp.395-416.
182. Procheş, Ş. and Cowling, R.M., 2006. Insect diversity in Cape fynbos and neighbouring South African vegetation. Global Ecology and Biogeography, 15(5), pp.445-451.
183. Prosdocimi, F., Peixoto, F.C. and Ortega, J.M., 2003. DNA sequences base calling by PHRED: Error pattern analysis.
184. Qian, W. and Zhang, J., 2008. Gene dosage and gene duplicability. Genetics, 179(4), pp.2319-2324.
185. Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F.O., 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41(D1), pp.D590-D596.
186. Randler, C., 2002. Avian hybridization, mixed pairing and female choice. Animal Behaviour, 63(1), pp.103-119.

187. Reed, K.M., Chaves, L.D., Garbe, J.R., Da, Y. and Harry, D.E., 2003. Allelic variation and genetic linkage of avian microsatellites in a new turkey population for genetic mapping. Cytogenetic and Genome Research, 102(1-4), pp.331-339.
188. Reynolds, C. and Cumming, G.S., 2015. The role of waterbirds in the dispersal of freshwater cladocera and bryozoa in southern Africa. African Zoology, 50(4), pp.307-311.
189. Reynolds, C., 2016. The role of waterbirds in the dispersal of aquatic organisms in southern Africa. PhD thesis, University of Cape Town, South Africa.
190. Rhoads, A. and Au, K.F., 2015. PacBio sequencing and its applications. Genomics, Proteomics and Bioinformatics, 13(5), pp.278-289.
191. Rhymer JM, 2006. S33-4 Extinction by hybridization and introgression in anatine ducks. In: Schodde, R. (ed.). Proceedings of the 23rd International Ornithological Congress Beijing. Acta Zoologica Sinica, 52 (Suppl.), pp.583–585.
192. Rhymer, J.M. and Simberloff, D., 1996. Extinction by hybridization and introgression. Annual Review of Ecology and Systematics, 27(1), pp.83-109.
193. Rhymer, J.M., Williams, M.J. and Braun, M.J., 1994. Mitochondrial analysis of gene flow between New Zealand mallards (*Anas platyrhynchos*) and grey ducks (*A. superciliosa*). The Auk, 111(4), pp.970-978.
194. Rizzi, R., Beretta, S., Patterson, M., Pirola, Y., Previtali, M., Della Vedova, G. and Bonizzoni, P., 2019. Overlap graphs and de Bruijn graphs: data structures for de novo genome assembly in the big data era. Quantitative Biology, 7(4), pp.278-292.
195. Roberts, M., Hayes, W., Hunt, B.R., Mount, S.M. and Yorke, J.A., 2004. Reducing storage requirements for biological sequence comparison. Bioinformatics, 20(18), pp.3363-3369.
196. Rocke, T. and Bollinger, T., 2007. Avian botulism. In: N. Thomas, D. Hunter and C. Atkinson, eds., Infectious Diseases of Wild Birds. Melbourne, Australia: Blackwell Publishing, pp.377-416.
197. Ruan, J. and Li, H., 2020. Fast and accurate long-read assembly with wtdbg2. Nature Methods, 17(2), pp.155-158.
198. Saalbach, K.P., 2022. Gain-of-function research. Advances in Applied Microbiology, 120, pp.79-111.

199. Sánchez-Pulido, L., Rojas, A.M., Van Wely, K.H., Martinez-A, C. and Valencia, A., 2004. SPOC: a widely distributed domain associated with cancer, apoptosis and transcription. BMC bioinformatics, 5, pp.1-6.

200. Sanderson, B.A., Araki, N., Lilley, J.L., Guerrero, G. and Lewis, L.K., 2014. Modification of gel architecture and TBE/TAE buffer composition to minimize heating during agarose gel electrophoresis. Analytical Biochemistry, 454, pp.44-52.

201. Sarais, F., Rebl, H., Verleih, M., Ostermann, S., Krasnov, A., Köllner, B., Goldammer, T. and Rebl, A., 2020. Characterisation of the teleostean κB-Ras family: The two members NKIRAS1 and NKIRAS2 from rainbow trout influence the activity of NF-κB in opposite ways. Fish & Shellfish Immunology, 106, pp.1004-1013.

202. Sathyanarayanan, A., Manda, S., Poojary, M. and Nagaraj, S.H., 2018. Exome sequencing data analysis.

203. Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O. and Thompson, J.D., 2020. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC genomics, 21(1), pp.1-20.

204. Seemann, T., 2022. Barrnap. [online] Available at: https://github.com/tseemann/barrnap/blob/master/README.md [Accessed: 2 June 2022].

205. Sekercioglu, C.H., 2006. Increasing awareness of avian ecological function. Trends in Ecology and Evolution, 21(8), pp.464-471.

206. Serjeantson, D., 2014. Duck: Domestication. In: C. Smith, ed. Encyclopedia of Global Archaeology. Springer, New York

207. Seutin, G., White, B.N. and Boag, P.T., 1991. Preservation of avian blood and tissue samples for DNA analyses. Canadian Journal of Zoology, 69(1), pp.82-90.

208. Shapiro, J.A. and von Sternberg, R., 2005. Why repetitive DNA is essential to genome function. Biological Reviews, 80(2), pp.227-250.

209. Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H., 2017. DNA sequencing at 40: past, present and future. Nature, 550(7676), pp.345-353.

210. Sim, S.B., Corpuz, R.L., Simmonds, T.J. and Geib, S.M., 2022. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in

PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics, 23(1), pp.1-7.

211. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), pp.3210-3212.

212. Sivic, J. and Zisserman, A., 2003. Video Google: A text retrieval approach to object matching in videos. In Computer Vision, IEEE International Conference on, 3, pp. 1470-1470.

213. Slatko, B.E., Gardner, A.F. and Ausubel, F.M., 2018. Overview of next-generation sequencing technologies. Current Protocols in Molecular Biology, 122(1), p.e59.

214. Smit, A., Hubley, R. and Green, P., 2013. RepeatMasker Open-4.0. [online] Available at: https://www.repeatmasker.org [Accessed: 22 May 2022].

215. Southern African Bird Atlas Project 2, (2022). [online] Available at: http://datazone.birdlife.org [Accessed: 04 February 2022].

216. Sović, I., Kronenberg, Z., Dunn, C., Barnett, D., Kingan, S., Drake, J. and Korlach, J., 2020. [Online PowerPoint presentation] Improved Phased Assembly using HiFi Data. SMRT Leiden 2020, Developer Conference. 27 April. Available at: https://www.pacb.com/videos/smrt-leiden-improved-phased-assembly-using-hifi-data/ [Accessed: 6 July 2022]

217. Stamatakis, A. and Aberer, A.J., 2013, May. Novel parallelization schemes for large-scale likelihood-based phylogenetic inference. In 2013 IEEE 27th International Symposium on Parallel and Distributed Processing (pp. 1195-1204). IEEE

218. Stanke, M. and Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, 19(suppl_2), pp.ii215-ii225.

219. Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D., 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics, 24(5), pp.637-644.

220. Stanke, M., Schöffmann, O., Morgenstern, B. and Waack, S., 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics, 7(1), pp.1-11.

221. Stavenga, D.G., Van Der Kooi, C.J. and Wilts, B.D., 2017. Structural coloured feathers of mallards act by simple multilayer photonics. Journal of The Royal Society Interface, 14(133), p.20170407.
222. Stephens, K., Measey, J., Reynolds, C. and Le Roux, J.J., 2019. Occurrence and extent of hybridisation between the invasive Mallard Duck and native, Yellow-billed Duck in South Africa. Biological Invasions, 22(2), pp.693-707.
223. Suri, J., Anderson, P.M., Charles-Dominique, T., Hellard, E. and Cumming, G.S., 2017. More than just a corridor: A suburban river catchment enhances bird functional diversity. Landscape and Urban Planning, 157, pp.331-342.
224. Szymanski, M., Barciszewska, M.Z., Erdmann, V.A. and Barciszewski, J., 2002. 5S ribosomal RNA database. Nucleic Acids Research, 30(1), pp.176-178.
225. Tiselius, A., 1937. A new apparatus for electrophoretic analysis of colloidal mixtures. Transactions of the Faraday Society, 33, pp.524-531.
226. Todesco, M., Pascual, M.A., Owens, G.L., Ostevik, K.L., Moyers, B.T., Hübner, S., Heredia, S.M., Hahn, M.A., Caseys, C., Bock, D.G. and Rieseberg, L.H., 2016. Hybridization and extinction. Evolutionary Applications, 9(7), pp.892-908.
227. Tomasiks, P. and Warren, D.M., 1996. The use of *Daphnia* in studies of metal pollution of aquatic systems. Environmental Reviews, 4(1), pp.25-64.
228. Tshabalala, N, Department of Forestry, Fisheries and the Environment of the Republic of South Africa, 2022. Agreement on the Conservation of African-Eurasian Migratory Waterbirds (The Hague, 1995): Report on the implementation of AEWA for the period 2018-2020
229. van Leeuwen, C.H., Van der Velde, G., van Groenendael, J.M. and Klaassen, M., 2012. Gut travellers: internal dispersal of aquatic organisms by waterfowl. Journal of Biogeography, 39(11), pp.2031-2040.
230. Vaser, R., Sović, I., Nagarajan, N. and Šikić, M., 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research, 27(5), pp.737-746.
231. Vieira, M.L.C., Santini, L., Diniz, A.L. and Munhoz, C.D.F., 2016. Microsatellite markers: what they mean and why they are so useful. Genetics and Molecular Biology, 39(3), pp.312-328.

232. Walsh, N., Dale, J., McGraw, K.J., Pointer, M.A. and Mundy, N.I., 2012. Candidate genes for carotenoid coloration in vertebrates and their expression profiles in the carotenoid-containing plumage and bill of a wild bird. Proceedings of the Royal Society B: Biological Sciences, 279(1726), pp.58-66.

233. Wang, J., Yang, W., Zhang, S., Hu, H., Yuan, Y., Dong, J., Chen, L., Ma, Y., Yang, T., Zhou, L. and Chen, J., 2023. A pangenome analysis pipeline provides insights into functional gene identification in rice. Genome Biology, 24(1), pp.1-22.

234. Wang, L., Guo, J., Xi, Y., Ma, S., Li, Y., He, H., Wang, J., Han, C., Bai, L., Mustafa, A. and Liu, H., 2020. Understanding the genetic domestication history of the Jianchang duck by genotyping and sequencing of genomic genes under selection. G3: Genes, Genomes, Genetics, 10(5), pp.1469-1476.

235. Waters, P.D., Patel, H.R., Ruiz-Herrera, A., Álvarez-González, L., Lister, N.C., Simakov, O., Ezaz, T., Kaur, P., Frere, C., Grützner, F. and Georges, A., 2021. Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. Proceedings of the National Academy of Sciences, 118(45), p.e2112494118.

236. Weissensteiner, M.H. and Suh, A., 2019. Repetitive DNA: The dark matter of avian genomics. In Avian Genomics in Ecology and Evolution. Springer, Cham, pp. 93-150.

237. Wheeler, T.J. and Eddy, S.R., 2013. nhmmer: DNA homology search with profile HMMs. Bioinformatics, 29(19), pp.2487-2489.

238. Whelan, S. and Goldman, N., 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Molecular biology and evolution, 18(5), pp.691-699.

239. Williams, M. and Basse, B., 2006. Indigenous gray ducks, *Anas superciliosa*, and introduced mallards, *A. platyrhynchos*. New Zealand: processes and outcome of a deliberate encounter. Acta Zoologica Sinica, 52, pp.579-582.

240. Wingfield, J.C., Ball, G.F., Dufty, A.M., Hegner, R.E. and Ramenofsky, M., 1987. Testosterone and aggression in birds. American Scientist, 75(6), pp.602-608.

241. Woodman, H.M., 1944. Notes on the Palaearctic Migrant and Resident Duck and Geese of Abyssinia. Ibis, 86(2), pp.151-166.

242. Woolson, R.F., 2007. Wilcoxon signed-rank test. Wiley encyclopedia of clinical trials, pp.1-3.

243. Wright, D.R. and Lee, A.T.K. 2017. Bird -friendly Habitat Management Guidelines for the endemic birds of the Fynbos Biome. BirdLife South Africa, Johannesburg.

244. Young, H.G., 1999. Comparative study of the courtship displays of Meller's Duck Anas melleri, Yellowbilled Duck *A. undulata* and Northern Mallard *A. platyrhynchos*. Ostrich, 70(2), pp.117-122.

245. Zhang, G., Li, B., Li, C., Gilbert, M.T.P., Jarvis, E.D., Wang, J. and Avian Genome Consortium, 2014a. Comparative genomic data of the Avian Phylogenomics Project. GigaScience, 3(1), pp.2047-217X.

246. Zhang, G., Li, C., Li, Q., Li, B., Larkin, D.M., Lee, C., Storz, J.F., Antunes, A., Greenwold, M.J., Meredith, R.W. and Ödeen, A., 2014b. Comparative genomics reveals insights into avian genome evolution and adaptation. Science, 346(6215), pp.1311-1320.

247. Zhang, Z., Jia, Y., Almeida, P., Mank, J.E., van Tuinen, M., Wang, Q., Jiang, Z., Chen, Y., Zhan, K., Hou, S. and Zhou, Z., 2018. Whole-genome resequencing reveals signatures of selection and timing of duck domestication. Gigascience, 7(4), p.giy027.

248. Zhong, M., Hansen, B., Nesnidal, M., Golombek, A., Halanych, K.M. and Struck, T.H., 2011. Detecting the symplesiomorphy trap: a multigene phylogenetic analysis of terebelliform annelids. BMC evolutionary biology, 11, pp.1-15.

Appendix

Table A1: Experimental data summary of the Yellow-billed Duck and comparative genomes

Assembly No.	Assembly	Organism	Common name	Organism code	Sex	Sequencing technology	Coverage
1	**Final YBD**	***A. undulata***	**Yellow-billed Duck**	***A. undulata***	**F**	**PacBio CLR and HiFi**	**16×**
2	GCA_002224875	*A. zonorhyncha*	Eastern Spot-billed Duck	*A. zonorhyncha*	F	Illumina HiSeq 2000	220×
3	GCA_015476345	*A. platyrhynchos*	Pekin Duck	Pekin-Pooled	Pooled M/F	PacBio RSII; PacBio Sequel	143×
4	GCA_003850225	*A. platyrhynchos*	Pekin Duck	Pekin-2015	No data	PacBio; BioNano; Hi-C	50×
5	GCA_002743455	*A. platyrhynchos*	Pekin Duck	Pekin-Male	M	Illumina HiSeq; 10X Genomics; Bionano	150×
6	GCA_017639285	*A. platyrhynchos*	Pekin Duck	CAU-Pekin	F	PacBio Sequel	93×
7	GCA_017639305	*A. platyrhynchos*	Laying Duck	CAU-Laying	F	PacBio Sequel	93×
8	GCA_008746955	*A. platyrhynchos*	Wild Mallard	CAU-Wild	F	PacBio Sequel	100×
9	GCA_002224895	*A. platyrhynchos*	Wild Mallard	D2L-Wild	F	Illumina HiSeq 2000	208×
10	GCA_900411745	*A. platyrhynchos*	Wild × domestic Mallard cross	Wild×Dom	No data	Illumina HiSeq 2000; HiSeq X Ten machine	64×
11	GCA_000355885	*A. platyrhynchos*	Pekin Duck	Pekin-Female	F	Illumina Genome Analyser	60×
12	GCA_016699485	*G. gallus* subp. *domesticus*	Chicken	*G. gallus*	F	PacBio CLR	102×

Yellow-billed Duck experimental data in bold.