

https://dspace.org/

Genome sequencing of the Southern Ground Hornbill (Bucorvus leadbeateri)

Patel, Jasmin Bharatkumar

2024-10

University of the Witwatersrand, Johannesburg

https://hdl.handle.net/10539/45111

Patel, Jasmin Bharatkumar. (2024). Genome sequencing of the Southern Ground Hornbill (Bucorvus leadbeateri). [PhD thesis, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/45111

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



Genome sequencing of the Southern Ground Hornbill (*Bucorvus leadbeateri*)

by

Jasmin Bharatkumar Patel

(1119140)

Thesis

Submitted in fulfilment of the requirements for the degree

Doctor of Philosophy

in

Molecular and Cell Biology

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

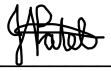
Supervisor: Professor Pieter De Maayer

Co-supervisor: Dr Jean Mollett

October 2024

DECLARATION

I declare that this thesis is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other University.



(Signature of candidate)

10th day of October 2024 at the University of the Witwatersrand, Johannesburg

PREFACE

The Southern Ground Hornbill (SGH – *Bucorvus leadbeateri*) is one of the largest hornbill species worldwide, known for its complex social structures and breeding behaviours. These birds, endemic to Africa, have been of great concern due to their declining populations and disappearance from historic ranges. Despite being the focus of numerous conservation efforts, with research forming an integral part of these initiatives, there is a lack of knowledge regarding the molecular biology aspects of this bird species. This study bridges the gap by presenting the first whole genome sequence of the SGH. The SGH genome was further explored using comparative genomics, genetic variant, and selection analysis, providing deeper insights into the evolution and adaptation of this species.

Chapter 1 comprehensively reviews pertinent literature on various aspects of avian evolution, including the role genomics has played in elucidating how these species have adapted and evolved. Furthermore, the current body of knowledge on the SGH is explored.

In Chapter 2 the entire genome sequence of the SGH was sequenced using Illumina short-read and Pacific BioSciences long-read datasets. Subsequently, the performance of various assembly approaches was evaluated to attain a high-quality assembly of the SGH. This was coupled with parameter optimisation and reference-based refinement to improve the SGH draft genome assembly. The final draft genome assembly was structurally annotated, providing insight into the genetic blueprint underpinning the SGH.

Chapter 3 presents the comparative genomic analysis of the SGH with the genomes of available hornbill species from the genera *Bucorvus* (*Bucorvus abyssinicus* and SGH) and *Buceros* (*Buceros bicornis* and *Buceros rhinoceros* subsp. *silvestris*). This included analysis of the pan-genome of the hornbill species, functional characterisation of the core and genus-specific elements of the pan-genome and analysis of orthogroups with evidence of paralogy.

In Chapter 4, a species-level comparative genomic analysis of the SGH and the Abyssinian Ground Hornbill (AGH) was performed. Here differences in the species-specific proteome of the two species were analysed and the functional implications of these differences on the adaptation and survival of these species were evaluated. Furthermore, genetic variations between the SGH and AGH were identified and selection analysis of key protein-coding genes

with high-impact variants was undertaken. This provided insight into the genetic diversity between the SGH and AGH.

Finally, the implications of the study on the understanding of the genetic basis underlying the evolution and adaptation of the SGH is discussed and the future perspective of large-scale population genetic studies is provided.

RESEARCH OUTPUTS

Journal articles

Patel, J., Mollet, J. and De Maayer, P., 2024. Whole genome sequencing, assembly and annotation of the Southern Ground Hornbill – *Bucorvus leadbeateri*. *Scientific data*. Article submitted on 26 September 2024, Submission ID: SDATA-24-03103

Patel, J., Grab, S. and De Maayer, P., 2023. Distinct microbial communities across a climatically versatile summit in the Lesotho highlands. *Ecology and Evolution*, *13*(3), p.e9891.

Hart, B., Patel, J., De Maayer, P., Nweke, E.E. and Bizos, D., 2023. Metataxonomic Analysis Demonstrates a Shift in Duodenal Microbiota in Patients with Obstructive Jaundice. *Microorganisms*, *11*(6), p.1611.

Conference outputs

Patel J., Mollett, J. and De Maayer P., 2024. Whole genome sequencing and assembly of the Southern Ground Hornbill (*Bucorvus leadbeateri*). BIO2024: LIFE – Air, Land and Water. Hybrid Event – South African Genetics Society and the South African Society for Bioinformatics. 23 September – 25 September 2024. Future Africa conference centre, Pretoria: Poster

Patel J., De Maayer P., Mollett J. and Botes A., 2023. Whole genome sequencing and assembly of the Southern Ground Hornbill (*Bucorvus leadbeateri*). The 14th Wits Cross-Faculty Postgraduate Symposium. 6-8 September 2023. University of the Witwatersrand, Johannesburg: Oral presentation

Patel J., De Maayer P., Mollett J. and Botes A. 2022. Unravelling the genome of the Southern Ground Hornbill (*Bucorvus leadbeateri*). International Ornithological Congress 2022 conference. 15 August – 20 August 2022. 28th IOCongress®, virtual format: Poster

Patel J., De Maayer P., Mollett J. and Botes A. 2022. Unravelling the genome of the Southern Ground Hornbill (*Bucorvus leadbeateri*). BIO2022: Bioscience, Big Data and the Fourth Industrial Revolution. Hybrid Event – South African Genetics Society and the South African Society for Bioinformatics. 24 April – 27 April 2022. STIAS Conference Centre, Stellenbosch: Poster

Patel, J., Grab, S. and De Maayer, P., 2021. Analysis of the microbial diversity across a mountain with distinct climatic conditions in the Lesotho highlands. South African Society for Microbiology (SASM) conference. 4 May – 6 May 2021. SASM Virtual conference: Oral presentation

DATA AVAILABILITY

The SGH genome sequence and assembly statistics have been submitted to DDBJ/ENA/GenBank under the BioProject ID PRJNA1116058 and BioSample accession SAMN41519907 and can be accessed using the GenBank accession number GCA_042477495.1. Additionally, the genome assembly has been successfully uploaded to Figshare and can be accessed via the following link: https://figshare.com/s/1f328465eb3495e540f4.

To my beloved family and furry friends. Your unwavering support has been my guiding light, illuminating my path when all was dim. Without you all in my life, I would not be where I am today.

ACKNOWLEDGEMENTS

I would like to thank the National Research Foundation (NRF) and the University of the Witwatersrand Postgraduate Merit Award (PMA) for their generous financial support throughout my PhD journey. Their backing has been instrumental in the progress and success of my research.

My sincere appreciation goes to Birdlife South Africa, the South African National Biodiversity Institute (SANBI), and the Mabula Ground Hornbill Project. Their financial contributions have made the sequencing of the Southern Ground Hornbill (SGH) genome possible, a critical component of this study.

I am profoundly grateful to Dr Kemp, whose guidance was invaluable in obtaining the necessary sample from a female SGH. Without her assistance, this project would not have been possible. I also extend my gratitude to Dr Katja Koeppel and the dedicated staff at Monte Casino Bird Gardens for their assistance and hospitality during the blood sample extraction process.

I owe a debt of gratitude to Professor Pieter De Maayer, whose unwavering support, supervision, and guidance have been a beacon throughout this research. His constructive feedback has been a guiding light in times of doubt, and his encouragement has given me the strength to persevere. I am also deeply thankful to Dr Jean Mollett for her consistent assistance and advice. Her motivation and readiness to lend a helping hand have been a source of inspiration.

I would like to acknowledge my colleagues and laboratory members for their support and assistance. Their camaraderie and cooperation have made this journey less daunting.

My heartfelt thanks go to my family and friends. Their continuous support, love, and guidance have been my rock throughout my life. Climbing this mountain would not have been possible without them. This achievement is as much theirs as it is mine.

Lastly, I would like to express my profound gratitude to Sai Baba. His teachings of “Shraddha aur Saburi” have been my guiding light in times of uncertainty. His blessings have been instrumental in helping me overcome numerous challenges in my life and throughout this PhD journey. His divine guidance has been a source of strength and resilience, enabling me to navigate the complexities of this journey with grace and perseverance.

TABLE OF CONTENTS

DECLARATION ii

PREFACE iii

RESEARCH OUTPUTS v

Journal articles v

Conference outputs v

DATA AVAILABILITY vii

DEDICATION viii

ACKNOWLEDGEMENTS ix

TABLE OF CONTENTS x

LIST OF FIGURES xiv

LIST OF TABLES xv

Chapter one – Avian genomics and its evolutionary significance 1

1.1 Introduction 2

1.2 Exploring the evolution and taxonomic classification of avian species 3

1.2.1 The evolutionary origins of avian species 3

1.2.2 Evolution of avian species over time 3

1.2.3 Taxonomic classification of avian species 5

1.3 The genomic architecture of avian species 5

1.3.1 Genome size 5

1.3.2 Chromosomes 7

1.3.3 Mitochondrial genome 9

1.3.4 Coding and non-coding genes 10

1.4 Resolving avian genomes through genome sequencing 12

1.5 Taxonomic classification and characteristics of hornbills 15

1.6 The Southern Ground Hornbill 16

1.6.1 Southern Ground Hornbill traits and distribution 16

1.6.2 The intricate breeding activity and social structure of the Southern Ground Hornbill 18

1.6.3 Conservation of the Southern Ground Hornbill 20

1.7 Rationale of study 21

1.8 References 22

Chapter two – Whole genome sequencing and assembly of the Southern Ground Hornbill (*Bucorvus leadbeateri*) 43

2.1 Abstract 44

2.2 Introduction 44

2.3 Methods and materials 46

2.3.1 Sample collection and DNA extraction 46

2.3.2 Library preparation and sequencing 47

2.3.3 Raw data quality control and processing 48

2.3.4 *De novo* genome assembly 48

2.3.5 Refining the draft SGH genome assembly 52

2.3.6 Genome annotation 54

2.4 Results 56

2.4.1 Raw sequence data assessment 56

2.4.2 *De novo* genome assemblies 57

2.4.3 Refinement of the draft SGH genome assembly 65

2.4.4 Comparison of the final SGH assembly and AGH assembly 68

2.4.5 Genome annotations 71

2.5 Discussion 73

2.6 Conclusion 76

2.7 References 77

Chapter three – Comparative genomic analysis of the Southern Ground Hornbill 90

3.1 Abstract 91

3.2 Introduction 91

3.3 Methods and materials 92

3.3.1 Genome sequences, structural and function annotation 92

3.3.2 Identification of orthologues and pan-genome analysis 93

3.3.3 Gene duplication events 94

3.3.4 Core genome phylogeny 95

3.3.5 Comparative analysis of hornbill mitochondrial genomes 95

3.4 Results 96

3.4.1 Hornbill genome metrices 96

3.4.2 Structural analysis of the hornbill genomes 98

3.4.3 Pan-genome analysis of the hornbills 101

3.4.4 Functional analysis of the accessory genome using KEGG pathways 103

3.4.5 Gene duplication events 116

3.4.6 Comparative analysis of the hornbill mitochondrial genomes 118

3.5 Discussion 123

3.5.1 The *Bucorvus* species exhibit greater genome sizes and genome complexity 123

3.5.2 The *Bucorvus* species genomes code for more genus-specific proteins 124

3.5.3 The *Bucorvus* and *Buceros* species evolved different repertoires of genus-specific proteins related to metabolism and organism systems 126

3.5.4 The genomes of the *Bucorvus* species gained new genes through gene duplications 128

3.6 Conclusion 129

3.7 References 130

Chapter four – Comparative genomic analysis of the Southern Ground Hornbill and the Abyssinian Ground Hornbill – Genetic variants and selection analysis 152

4.1 Abstract 153

4.2 Introduction 153

4.3 Methods and materials 155

4.3.1 Comparison of the SGH and AGH unique and gene duplicated proteins 155

4.3.2 Analysis of genetic variations 155

4.3.3 Selection analysis 160

4.4 Results 160

4.4.1 Species-specific differences between the SGH and AGH proteomes 160

4.4.2 Gene duplications between the SGH and AGH 164

4.4.3 Analysis of genetic variations 166

4.4.4 Selection analysis 174

4.5 Discussion 175

4.5.1 The SGH and AGH evolved different repertoires of proteins related to metabolism and immune response 175

4.5.2 SVs in immune-related proteins could affect the fitness and survival of the SGH 177

4.5.3 Proteins crucial for survival and reproduction are conserved via negative selection 178

4.6 Conclusion 179

4.7 References 180

SUMMARY 198

APPENDIX 200

LIST OF FIGURES

Figure 1.1: Photographic examples of the SGH.. 17

Figure 2.1: Workflow of the SGH *de novo* assembly approach.....49

Figure 2.2: Identification of optimal k-mer length for the short-read Illumina dataset.. 58

Figure 2.3: Sequence homology of the unaligned SGH contigs assembled.. 70

Figure 2.4: Composition of repetitive elements identified within the SGH draft genome.. 71

Figure 2.5: tRNA carrying amino acids and rRNA abundance within the SGH draft genome.. 72

Figure 3.1: Assessment of genome assembly and annotation quality using BUSCO.. 98

Figure 3.2: Pan-genome analysis of the four hornbill species. 101

Figure 3.3: Relative proportion of the core and accessory genome of the four hornbill species in each COG super-functional group.. 102

Figure 3.4: Proteins involved in the glycerolipid metabolism pathway.. 105

Figure 3.5: Proteins involved in the oxidative phosphorylation pathway.. 107

Figure 3.6: Proteins involved in the insulin secretion pathway.. 111

Figure 3.7: Proteins involved in the spliceosome.. 114

Figure 3.8: Phylogenetic placement of the four hornbills and chicken based on their mitochondrial genomes and protein homology.. 120

Figure 3.9: The mitochondrial genome map of the hornbill species and the chicken mapped against the SGH as a reference.. 122

Figure 4.1: Workflow of the read-mapping based approach for variant calling within the SGH genome……...156

Figure 4.2: Relative proportion of core, species-specific and duplicated proteins in the SGH and AGH in each COG super-functional group.. 162

Figure 4.3: Genomic variants identified between the SGH and AGH are classified according to the a) region in the gene and b) the impact of the variant on the protein-coding sequence. .171

Figure 4.4: Relative proportion of proteins with genomic variants between the SGH and AGH in each COG super-functional group.. 173

LIST OF TABLES

Table 2.1: Genome sequencing results of the PacBio and Illumina reads 57

Table 2.2: ABySS assembly statistics across a range of k-mer lengths evaluated.....60

Table 2.3: Canu and wtdbg2 long-read assembly statistics compared to optimal short-read assembly.....61

Table 2.4: Assembly statistics of the SGH draft genome assembly per improvement stage ...64

Table 2.5: Comparison of the SGH assembly and the AGH assembly 69

Table 3.1: Structural annotation statistics of the hornbill genomes……..97

Table 3.2: Relative proportion of repetitive elements within the hornbill species 100

Table 3.3: Annotation statistics of the mitochondrial genomes of the hornbill species and the chicken 119

Table 3.4: Sequence similarity matrix for the mitochondrial genomes of the hornbill species and the chicken 121

Table 4.1: Number of variants detected from the Illumina short reads and PacBio long reads using a consensus-based approach……166

Table 4.2: Different SV types detected from the Illumina short reads and PacBio long reads pre- and post-filtering 168

Table 4.3: Annotation statistics of the genomic variants determined using SnpEFF.....169

Chapter one – Avian genomics and its evolutionary significance

1.1 Introduction

Aves is one of the most diverse classes of vertebrates that are characterised by feathers, toothless beaks, laying of hard-shelled amniotic eggs, high metabolic rates, a four-chambered heart and strong yet lightweight skeletons (Benton *et al*., 2015; Brusatte *et al*., 2015; Scanes, 2020; Wu and Wang, 2019). These species inhabit a wide range of environments across the world, from the polar ice caps to tropical forests and deserts (Ainley *et al*., 2016; Dean and Williams, 2004; Liang *et al*., 2024). Their unique adaptations, which are key to their survival, are as diverse as the environments they inhabit. A common example of this is the Darwin finches that display extensive variations in beak morphology as an adaptive response to their environment (Lamichhaney *et al*., 2016, 2015). With ~ 10,980 living and 159 extinct bird species, they present a vast array of sizes, colours and characteristics, from the tiny hummingbird to the majestic ostrich (Brusatte *et al*., 2015). Each individual of this class plays vital roles in their respective environments such as seed dispersal, scavenging, controlling invertebrate populations, pollination and predation (Whelan *et al*., 2008). Furthermore, they are important indicators of ecosystem health, and their presence and behaviour can provide insight into environmental changes (Whelan *et al*., 2008).

Since the sequencing of the first avian genome (a female red junglefowl *Gallus gallus*, the wild ancestor to the domestic chicken *Gallus domesticus*) in 2004 (Hillier *et al*., 2004), initiatives such as the Avian Phylogenomic and Bird 10,000 Genomes (B10K) projects have resulted in a veritable explosion in available bird genome sequences (Jarvis *et al*., 2015; Zhang, 2015). As such, the genome sequences and assemblies of major representatives from all avian orders and 236 avian families are available, with major progress being made in obtaining the genomes of all bird taxa at the genus and species level (B10K, 2024). These genomes have provided great insight into the evolutionary history of avian species and the genetic basis underlying key characteristics such as flight, beak morphology and vocalisation patterns (Jarvis *et al*., 2015; Zhang *et al*., 2014).

While avian orders such as Passeriformes (perching birds) have received substantial genome coverage due to their large ecological diversity (Jarvis *et al*., 2014; Mackiewicz *et al*., 2019; Matsunaga and Okanoya, 2009; Stiller *et al*., 2024), other orders, such as the Bucerotiformes (hornbills and hoopoes), have been poorly represented with only three whole genome sequences available. This order incorporates the Southern Ground Hornbill (*Bucorvus leadbeateri*), a species of major relevance in the African savannah biome and the focus of this

study. The future of this species is of great concern due to its dwindling populations. The Southern Ground Hornbill (SGH) has been a part of several conservation efforts which integrate a large research component (Jordan, 2011; Morrison *et al.*, 2005); however, this has mainly focused on the bird's survival, lifestyle, and behavioural patterns. Genetic information pertaining to this species remains elusive. This study focused on 1) the resolution of the whole genome sequence of the SGH and 2) the performance comparative genomic analysis with the available genomes of other hornbill species, to gain insight into the genetic basis underpinning the biology, ecology, behaviour and evolution of the SGH. Here literature pertinent to key aspects of the study will be discussed.

1.2 Exploring the evolution and taxonomic classification of avian species

1.2.1 The evolutionary origins of avian species

Birds evolved from small theropod dinosaurs over 150 million years ago during the Middle to Late Jurassic (Zhang *et al*., 2014). Archaeological evidence recovered from the Late Jurassic Solnhofen Limestone in Bavaria, Germany, in 1861 suggested *Archaeopteryx* as the first ancestor of avian species (Dodson, 2000). Fossil records of *Archaeopteryx* indicated that it was a medium-sized bird similar to a crow, with a long neck that was flexible and a short and compact body (Marshall, 2013). Furthermore, the presence of teeth impaled within the sockets of the jaw was noted together with bird-like features such as feathers on the vertebrae and reptilian features such as a long tail, amphicoelous vertebrae, a fibula as long as the tibia and free metacarpals and metatarsals (Marshall, 2013; Padian and Chiappe, 1998). Subsequently, many more fossils of birds dating back to the Cretaceous period have been found in places such as China, Madagascar, Mongolia and Spain (Padian and Chiappe, 1998). These findings have expanded the understanding of avian evolution well beyond *Archaeopteryx*, delineating a more comprehensive path for the evolution of prehistoric avian species to modern birds. This broader perspective incorporates a diverse array of early avian ancestors, providing a richer context for understanding the evolutionary transitions that have shaped the bird species as seen today.

1.2.2 Evolution of avian species over time

Life on Earth has gone through substantial change due to events of periodic mass extinctions. Mass extinction caused the global death of animal, plant and microorganism species, while those that were able to survive underwent rapid evolution (Alters, 2000). The emergence of

new habitats and environmental niches allowed for adaptive radiation and speciation, driving diversification within avian species due to habitat segregation (Alters, 2000). Evidence of diversification in birds was first recorded in fossils from Jehol, located in the western Liaoning Province, and neighbouring areas of northeastern China (Zhou, 2014). These fossils date back to over 120 million years ago and are some of the oldest avian fossils occurring after *Archaeopteryx* (Brusatte *et al*., 2015; Zhou and Zhang, 2007). Despite being some of the most diverse species for their time, avian species from the Early Cretaceous period only account for half of the global diversity recorded in Mesozoic birds and even less compared to modern-day avian species (Zhou and Zhang, 2007).

With birds in particular, the adaptation for flight became more pronounced compared to their dinosaurian ancestor, including groups such as Avialae and Enantiornithes (O'Connor, 2022; Padian and Chiappe, 1998). Bones within the hand and hip gridle fused to increase the strength of the skeleton for flight, while the breastbone broadened and a keel developed down the middle of the chest for the attachment of the flight muscles (Gatesy and Baier, 2005; Padian and Chiappe, 1998). The alula (required for flight control at low speeds) and the long first toe (required for perching) evolved (Hutchinson, 2001; Padian and Chiappe, 1998). Furthermore, the reptilian-like tail shortened into a tail stump, the skull and vertebrae bones lightened for flight and the impaled teeth disappeared, resulting in an edentulous (toothless) beak (Bhullar *et al*., 2016; Gatesy and Dial, 1996; Padian and Chiappe, 1998). This gave rise to several of the features identified in the modern-day birds. However, the mass extinction that took place at the end of the Mesozoic era during the Cretaceous period 65 million years ago led to the disappearance of dinosaurs and at least three-quarters of all plant and animal species (Whitehouse, 2018). Many avian species from major groups that evolved during the Cretaceous period also died out during this period (Padian and Chiappe, 1998). Those that survived the Cretaceous mass extinction contributed to the subsequent evolution and diversification of many present-day avian species (Brusatte *et al*., 2015; Yan *et al*., 2020).

At present, there is a broad variety of avian species with great diversity in size, colour and characteristics. For example, the hummingbird weighs approximately two grams, while the largest bird on Earth, the Ostrich, weighs ~140,000 grams (Brusatte *et al*., 2015). Birds have evolved to have bodies with several distinct characteristics, not limited to lightweight skeletons optimized for flight, high metabolic rates, highly encephalised brains and keen senses (Balanoff *et al*., 2013; Brusatte *et al*., 2015; Marugán-Lobón *et al*., 2016).

1.2.3 Taxonomic classification of avian species

At present, the class Aves is one of the most species-rich among the kingdom Animalia and phylum Chordata, with ~ 10,980 living and 159 extinct bird species known. Early classification placed birds, reptiles and mammals under a general group of amniotes on the basis of their tetrapod nature (a vertebrate with four limbs or those that evolved from four-limbed ancestors), and the shared feature of an amniotic egg composed of a shell and extraembryonic membranes (chorion, allantois and amnion) (Benton *et al*., 2015). These amniotic eggs were either laid on land or retained within the mother allowing for terrestrial survival (Benton *et al*., 2015; Rafferty and Reina, 2012). Despite sharing characteristics such as vocal communication, high rates of metabolism, a four-chambered heart, parental care and thermoregulation capabilities, birds separated from mammals over 318 million years ago (Scanes, 2020; Wu and Wang, 2019). Compared to mammals, birds are much more closely related to ancient reptiles such as lizards and snakes (Scanes, 2020; Sfetcu, 2014). Particularly, crocodiles, which belong to the order Crocodylia, are part of their sister group. Birds and crocodiles both fall under the clade Archosauria, which is a subset of the larger group Sauropsida (Scanes, 2020; Sfetcu, 2014).. Currently, avian species are divided into two super orders, Paleognathae, which mainly incorporate flightless birds, and Neognathae, which is highly diverse and comprises all other bird species (Scanes, 2020; Sfetcu, 2014). Avian species in both super orders are further classified into 253 families and 2,385 genera (Donsker and Rasmussen, 2023). Given the extensive morphological and physiological changes observed in avian species throughout their evolution, how are these changes reflected at the genomic level?

1.3 The genomic architecture of avian species

1.3.1 Genome size

The genomes of avian species are typically small and show relatively low amounts of variation in genome size compared to other tetrapod vertebrates (Kapusta *et al*., 2017). Cytological investigations of 897 avian species indicated that the genome sizes of avian species range between 0.9 Gigabase pairs (Gb) for the Black-chinned Hummingbird (*Archilochus alexandri*) to 2.1 Gb for the common Ostrich (*Struthio camelus*) (Gregory *et al*., 2007). In comparison, mammalian genome sizes (810 evaluated genomes) show a broader range, between 1.6 Gb for the Carriker's round-eared bat (*Lophostoma carrikeri*) to 8.2 Gb for the Red viscacha rat (*Tympanoctomys barrerae*). Humans have a genome size of 3.4 Gb, which is substantially

larger than the largest avian genome (Gregory *et al*., 2007). Similarly, squamates, reptiles such as lizards and snakes, also display a large variation in genome size (227 species) (Saha *et al*., 2023). The genomes of these species can range between 1.05 Gb for the mionecton skink (*Chalcides mionecton*) and 3.93 Gb for the armadillo girdled lizard (*Ouroborus cataphractus*) (Saha *et al*., 2023). Amongst reptiles, the tuatara (*Sphenodon punctatus*) also has a large genome size of 4.3 Gb, which is ~ 2 times larger than the largest avian genome *(*Gemmell *et al*., 2020).

The evolution of the avian genome size is multifaceted and based on several different hypotheses. One strongly supported hypothesis is the adaptive evolution of the avian genome size based on its ability of flight (Bravo *et al*., 2021). Adaptive evolution occurs when an organism develops beneficial traits, through natural selection, that enhance its fitness for a particular environment (Hughes, 1999). The genome size is considered to be reduced as an adaptation mechanism for high metabolic demand that is required for powered flight (Hughes and Piontkivska, 2005). Furthermore, in mammals and avian species, there is a strong correlation between the genome and cell size of organisms (Hughes and Piontkivska, 2005; Szarski, 1976). Cells with a larger size have a lower surface area to volume ratio allowing for less gas exchange (Gregory, 2001). Thus, as avian species have a higher metabolic demand, their cells are typically smaller than those of mammals (Hughes and Piontkivska, 2005). This hypothesis is supported by the finding of the smallest avian genome belonging to the hummingbird, which has an energy-intensive hovering flight mechanism, compared to the ostrich which has the largest genome and is a non-flying bird (Gregory *et al*., 2009; Hughes and Hughes, 1995).

It has also been hypothesised that the reduced genome sizes in avian species are a result of neutral evolution, during which the loss of genomic DNA that occurred in the avian ancestors was due to genetic drift, which could explain why flightless avian species have larger genome sizes (Bravo *et al*., 2021; Hughes and Piontkivska, 2005). The accumulation of mutations and insertion of transposable elements (TEs) in flightless birds aligns with the theory of nearly-neutral evolution in smaller populations, where genetic drift can lead to the fixation of slightly deleterious mutations (Ohta, 1992). In contrast, a recent study by Kapusta and co-workers, showed that flightless avian species had lower deleterious mutation rates than flying avian species. Furthermore, the larger genome sizes in flightless birds did not always result from the

accumulation of mutations and TEs but rather represent a balance between the accumulation and loss of DNA during evolution (Kapusta *et al*., 2017).

1.3.2 Chromosomes

For over 100 million years the evolution of most avian karyotypes has remained relatively stable and conserved, with the exception of certain groups, such as parrots (Huang *et al*., 2022). The avian karyotype consists of a diploid number of approximately 40 chromosomes (2n = 80), a number that has remained relatively constant in most avian species (Christidis, 1990; Huang *et al*., 2022). One study compared the karyotypes of 723 avian species, of which 63% had a karyotype of 2n = 76 to 86 chromosomes, while 24% had a karyotype of 2n = 66 to 74 (Christidis, 1990). Extremes on the karyotype spectrum include the common kingfisher (*Alcedo atthis* - 2n = 132 or 138 chromosomes) and the grey go-away bird (*Corythaixoide concolor* - 2n = 136 to 142 chromosomes) at the higher end, and the stone curlew (*Burhinus oedicemus* - 2n = 40 chromosomes) and the trumpeter hornbill (*Ceratogymna buccinator* - 2n = 40 chromosomes) at the lower end (Christidis, 1990). In order to determine the chromosomal composition of the avian karyotype, Takagi and Sasaki (1974) compared the composition of the karyotypes of 48 avian species across 12 orders. It was found that a large proportion (~ 30 pairs) of the chromosomal pairs in avian species are microchromosomes, followed by macrochromosomes (~ 10 pairs) and one pair of sex chromosomes (Takagi and Sasaki, 1974).

Avian microchromosomes are typically < 20 Megabases (Mb) in length and are between 0.5 and 2.5 µm in size, whereas macrochromosomes are >20 Mb in length and are between 3 and 6 µm in size (Burt, 2002; Huang *et al*., 2022). Size is not the only distinguishable characteristic of micro- and macrochromosomes, their presence and evolution, as well as their composition within the avian genome also differ. While macrochromosomes are more comparable to mammalian chromosomes in size and constitute ~70% of the avian genome (Damas *et al*., 2019), microchromosomes are characteristics of avian and reptilian karyotypes (Burt, 2002). With the exception of the order Crocodilia, microchromosomes diverged from the vertebrate lineage over 680 Million years ago (Burt, 2002; Janke and Arnason, 1997; Waters *et al*., 2021). These chromosomes are thus representatives of an archaic linkage group of ancestral vertebrates, as they have completely disappeared from the current mammalian lineage (Liu *et al*., 2021; Waters *et al*., 2021).

Despite being relatively small to the point of being impossible to detect using classic cytogenetic methods, microchromosomes incorporate 50% of all avian genes (Smith *et al*., 2000; Waters *et al*., 2021). Compared to macrochromosomes they further have a higher guanine-cytosine (GC) content and a number of CpG islands, which are regions of the genome that are rich in cytosine and guanine dinucleotides (Auer *et al*., 1987; Janitz and Janitz, 2011; McQueen *et al*., 1998). Microchromosomes are considered to have gene-rich chromosomal domains as they replicate in the early stages of the S phase and are hyperacetylated, thereby playing a crucial role in the regulation of gene transcription (McQueen *et al*., 1998; Sterner and Berger, 2000). They have functional centromeres and telomeres that are stably maintained during the process of mitosis and meiosis and have a substantially higher recombination rate compared to the macrochromosomes (Burt, 2002; Krishan, 1964; Solovei *et al*., 1994).

Unlike mammalian sex chromosomes which are denoted by XY, with the female being homogametic (XX) and the male being heterogametic (XY), avian sex chromosomes are depicted by WZ with the male being homogametic (ZZ), while the female is heterogametic (ZW) (Delbridge and Graves, 1999; Stiglec *et al*., 2007). The avian sex chromosomes differentiated from a pair of autosomal chromosomes possibly over 100 million years ago due to the suppression of crossing-over between the W and Z chromosomes, likely due to inversions on the sex-specific W chromosome (Fridolfsson *et al*., 1998; Handley *et al*., 2004; Q. Zhou *et al*., 2014). With the lack of crossing-over, the sex-specific chromosome diverged from the Z chromosomes which retained 98% of the ancestral autosomes genes, while the W chromosomes experienced genetic degradation over time (Bellott *et al*., 2010).

The extent of divergence of the W and Z chromosomes varies across avian species (Bellott *et al*., 2010). The Z and W chromosomes are generally homomorphic (similar in morphology) in ratites, an avian clade containing the orders Rheiformes (rhea), Struthioniformes (ostrich), Casuariiformes (emus and cassowary) and Apterygiformes (kiwi) from the super order Paleoneognathae (Ishijima *et al*., 2014). They have a relatively conserved karyotype, with the Z and W chromosomes having a similar size and very low levels of differentiation (Damas *et al*., 2019; Torgasheva *et al*., 2021). These chromosomes also recombine across most of their lengths and thus have a greater rate of homology and synteny (Mank and Ellegren, 2007; Torgasheva *et al*., 2021). In contrast, in the super order Neognathae, and parts of the Paleoneognathae, these two sex chromosomes are heteromorphic (different in morphology) and are differentiated by size, with the W chromosome appearing much smaller than the Z

chromosome. The W chromosome also has a lower gene content rich in heterochromatic regions, indicating genetic decay, and has a greater composition of repeats (Damas *et al*., 2019). Recombination between the two chromosomes is also restricted, with the W chromosome possessing a very short pseudoautosomal region (PAR), which is a region of homology between the Z and W chromosome near the telomere that recombines (Smeds *et al*., 2014; Torgasheva *et al*., 2021).

Sex determination in avian species is controlled by the genes *doublesex* and *mab-3*-related transcription factor (*DMRT1*) linked to the Z chromosome and occurs in a dosage-dependent manner (Ishijima *et al*., 2014). *DMRT1* has been found in many avian species, including the chicken (*Gallus gallus*), and has major implications for testicular formation in males (Ishijima *et al*., 2014). A study using RNA interference to alter the expression of *DMRT1* in chickens highlighted the importance of this gene in male gonadal development (Smith *et al*., 2009). Its absence on the W chromosome and impact on female development indicated that the double dosage of *DMRT1* suppresses female development in chickens and gives rise to males (Smith *et al*., 2009). Additionally, sexual dimorphism, differences in morphological traits between sexes, in avian species can be influenced by other genes (Liu *et al*., 2024). These may include those involved in melanin production, which influences plumage coloration in avian species. For example, melanocortin receptor 1 (*MC1R*), microphthalmia-associated transcription factor (*MITF*), tyrosinase (*TYR*) gene and tyrosinase-related protein 1 (*TYRP1*) are associated with melanin synthesis and feather coloration, contributing to visible dimorphic traits in avian species (Liu *et al*., 2024).

1.3.3 Mitochondrial genome

The mitochondrial genome (mtDNA) is a small circular DNA molecule located within the mitochondrion of cells and is inherited from mother to offspring. The avian mtDNA is a double-stranded circular molecule that ranges between 16.3 to 23.1 kilobase pairs in size (Yuan *et al*., 2023). It harbours 37 genes (thirteen protein-coding genes, twenty-two transfer RNAs (tRNAs) and two ribosomal RNAs (rRNAs)) and one or two control regions (non-coding regions that play a regulatory role in transcription and replication). The structure of mtDNA is largely conserved amongst all vertebrates, however, the gene order of these elements varies substantially within these taxa (Boore, 1999; Desjardins and Morais, 1990; Yuan *et al*., 2023). Compared to the mtDNA gene order first characterised in the chicken, the mtDNA of several avian orders including the Bucerotiformes, Charadriiformes, Gruiformes, Passeriformes and

Pelecaniformes display different gene orders (Akiyama *et al*., 2017; Caparroz *et al*., 2018; Desjardins and Morais, 1990; Mindell *et al*., 1998; Sammler *et al*., 2011; Verkuil *et al*., 2010; Zhou *et al*., 2014). These differences arise through tandem duplications of the control region which results in multiple copies of certain segments. Over time certain copies of these duplicated segments may be randomly lost, resulting in different gene orders (Zhou *et al*., 2014).

mtDNA has played a vital role in improving our understanding of avian evolution and phylogenetics. Studies of avian mtDNA have assisted in resolving the phylogenetic relationships of major avian orders and provided a clear distinction between the super orders Paleognathae and Neognathae (Gibb *et al*., 2007; Slack *et al*., 2007a). It has also aided in estimating the time of origin and biogeographical evolution of these species from ancestral lineages in the Cretaceous period (Arcones *et al*., 2021; Gibb *et al*., 2007; Harrison *et al*., 2004; Slack *et al*., 2007, 2006). Sequencing of complete mtDNA of extinct avian species from museum samples has further expanded the historical genetic diversity and evolutionary relationships of these avian taxa (Anmarkrud and Lifjeld, 2017; Grealy *et al*., 2019). These collections thus serve as crucial reservoirs for studying long-term changes in genomic diversity and understanding the evolutionary processes underpinning ancestral and modern avian species (Grealy *et al*., 2019).

1.3.4 Coding and non-coding genes

Coding regions, that are translated into functional proteins, despite their vital role in many structural and functional processes within cells and organisms, only make up ~ 2 to 3 % of most avian and vertebrate genomes (Harrow *et al*., 2009; Jax *et al*., 2018). In avian species, however, protein-coding genes are on average 50 % and 27% shorter than mammalian and reptilian genes, respectively (Zhang *et al*., 2014), with a large number of these protein-coding genes further being lost (Lovell *et al*., 2014). A study comparing the genomes of 60 avian species to those of other vertebrate lineages indicated that approximately 274 protein-coding genes present within most vertebrate lineages were lost within the avian genomes studied (Lovell *et al*., 2014).

Non-coding DNA was long considered as ‘junk DNA’ with no function. However, with the increasing availability of genome sequences from a broader lineage of vertebrate taxa, the importance and functions of non-coding DNA have been brought to the fore (Barrett *et al*., 2012). Despite lacking the ability to produce functional gene products, non-coding DNA can

contribute up to 98% of the genome and plays a crucial role in the regulation of transcriptional, and post-transcriptional events, RNA processing, modification, splicing, translation and replication (Shabalina and Spiridonov, 2004; Simna and Han, 2022; Zhang and Lupski, 2015; Zhang *et al*., 2019). Non-coding DNA includes regions such as the 5′ and 3′ untranslated regions (UTR), introns, promoter and enhancer regions (Barrett *et al*., 2012; Shabalina and Spiridonov, 2004). Apart from these elements, a large proportion of non-coding DNA within genomes is comprised of repetitive elements (de Koning *et al*., 2011).

Repetitive elements are DNA sequences that occur in multiple copies within the genome (Mehrotra and Goyal, 2014). These include tandem repeats, such as satellites, and interspersed repeats, such as transposable elements (TEs) (Kordis, 2010). One of the first avian TEs to be characterised was the Chicken Repeat 1 (CR1) (Kapusta and Suh, 2017). CR1 is one of the most prevalent TEs present in avian species and can account for 39 to 88% of all TEs present within a genome. Currently, at least fourteen CR1 families have been identified, all of which have diversified from a common CR1 ancestor during the early evolution of avian species (Kapusta and Suh, 2017). Furthermore, six other ancient CR1 families were identified that were present in avian and crocodilian lineages, but have gone extinct over time (Kapusta and Suh, 2017). Many other repetitive elements have been identified in avian genomes that have provided a crucial backbone for several areas of biological research, including population and conservation genetics studies and to investigate the evolution of avian species (Abdul-Muneer, 2014).

Repetitive elements are crucial for creating genetic variation within populations through events such as genomic rearrangements, shuffling of genes and the regulation of gene expression (Mehrotra and Goyal, 2014; Oliveira *et al*., 2017). Furthermore, these elements play a major role in sex chromosome differentiation and the size of the sex chromosome W in many avian species, as observed in bird species from the orders Caprimulgiformes, Gruiformes, Passeriformes and Psittaciformes (Barcellos *et al*., 2019; de Oliveira Furo *et al*., 2017; de Souza *et al*., 2020; Gunski *et al*., 2019). The composition of repetitive elements varies greatly within genomes such that in avian species they account for only 4 to 10% of the genome, whereas within mammalian genomes, these elements can account for 40 to 50% of the genome (Zhang *et al*., 2014). A possible explanation for the variation and reduced repetitive content within avian genomes is a result of the adaptation of avian species to the high oxidative metabolism required for flight in many species (de Souza *et al*., 2020). The loss of repetitive elements,

large segments of DNA and genes over time may be linked to the substantial difference noted in both the composition and size of avian genomes compared to other vertebrates (de Souza *et al*., 2020; Kapusta *et al*., 2017). These variations in the composition and characteristics of avian genomes compared to other vertebrates have sparked interest in the scientific community, leading to the advent of avian genomics (Zhang *et al*., 2014). The field aims to explore the genomic origins and evolution of such differences and their implications through the use of advanced sequencing techniques (Wu *et al*., 2021).

1.4 Resolving avian genomes through genome sequencing

Avian genomics is a relatively new field, with the sequencing of the chicken genome in 2004 marking its birth (Hillier *et al*., 2004). The chicken genome was sequenced to bridge the evolutionary gap between mammals and other amniotes as a representative of the modern descendants of dinosaurs and non-mammalian amniotes, providing the first glance into the genetic make-up of an avian species (Hillier *et al*., 2004). This laid the foundation for subsequent genome sequencing projects, including the sequencing of other avian species such as the zebra finch and turkey in 2010 (Dalloul *et al*., 2010; Warren *et al*., 2010). Advancements in sequencing technologies further led to the introduction of large-scale genomics projects such as the Genome 10K Project established in 2009, which aimed to sequence 10,000 or more vertebrate genomes (Koepfli *et al*., 2015). The Genome 10K project soon gave rise to several sister projects.

Following the launch of the Avian Phylogenomic Project, the genomes of 48 avian species across most avian orders were sequenced and assembled. This provided an overview of the evolutionary history between avian species, the impact of evolution on avian genomes (size, composition, and sex chromosomes) and traits such as the molecular basis of powered flight, loss of teeth, advanced visualisation systems and vocalisation patterns (Jarvis *et al*., 2015). This project laid the foundation for the Bird 10K (B10K) project which aimed to sequence the genomes of all extant bird species, approximately 10,500 species (Zhang, 2015).

Phase I and II of this project have already resulted in the sequencing and assembly of the genomes of 363 avian species and 218 families (accounting for 92% of all avian families) by 2014 and 2020, respectively (B10K, 2024; Jarvis *et al*., 2014; Zhang, 2015; Zhang *et al*., 2014). As of April 2024, an additional 1,833 genomes have been sequenced and assembled at the genus level and 3,811 at the species level in Phase III and IV, respectively (B10K, 2024).

According to the International Union of Conservation of Nature (IUCN) and BirdLife International approximately 12.8% of all avian species globally are considered threatened, with 755 species classified as vulnerable and 423 endangered (BirdLife International, 2022). The B10K project is thus providing an understanding of numerous aspects of the biology, behaviour and evolution of avian species which in turn can provide important information to assist in the conservation of these species (Stiller and Zhang, 2019).

Genomics has provided insight into various aspects of life in avian species including beak morphology, dietary adaptations and vocalisation (Wu *et al*., 2021). Darwin's finches from the Galápagos Islands are one of the most iconic examples portraying variations in beak morphology. Genomic studies of 120 individuals representing all Darwin's finch species have revealed several candidate genes linked to craniofacial development and those that drive character displacement (morphological divergence in beak size and shape) during challenging conditions (Lamichhaney *et al*., 2016, 2015). The variation in these genes leading to variations in beak morphology forms the basis for Darwin's finches utilizing different food sources and therefore reducing competition between them (Lamichhaney *et al*., 2016, 2015). Similarly, genomic studies of ground tits identified multiple genes associated with osteoblast and osteoclast differentiation that play a role in beak development (Cheng *et al*., 2017). Furthermore, a comparative analysis of multiple ground tit genomes identified candidate genes associated with bone development and morphology (Cheng *et al*., 2020). These genes underpin the variation in beak morphology noted in these birds, allowing them to adapt to ground foraging and cavity nesting in high-altitude terrestrial regions (Cheng *et al*., 2020, 2017).

Genomic studies of hummingbirds have identified a set of genes associated with taste receptors and metabolism that allow them to thrive on a nectar-rich diet. In particular, the *TAS1R1* gene, that is linked to umami taste in vertebrates, has been found to undergo several variations to support a nectar-based diet in hummingbirds and its high energy requirements (Baldwin *et al*., 2014). Vultures have also received substantial attention due to the evolution of their specialised diet. These birds are obligate scavengers that feed on dead and decaying animal matter that harbour potentially pathogenic microorganisms (Zou *et al*., 2021). Genomic analysis has resulted in the identification of several genes associated with gastric acid production and immunity (Zou *et al*., 2021). Interestingly, several genes and microorganisms associated with gas gangrene and food poisoning were identified in these birds (Zepeda Mendoza *et al*., 2018).

These studies reveal the interplay between genetics and host-associated microbiota, enabling vultures to thrive on their harsh, bacteria-laden diet (Chen *et al*., 2023).

Genomic studies in avian species have also helped understand the genetic basis underlying avian vocalisation. Neurological studies have led to the identification of complex neural circuits forming the song system in songbirds, parrots and hummingbirds involved in vocal learning and sound production via imitation (Matsunaga and Okanoya, 2009). Genomic studies further assisted in revealing several candidate genes involved in vocal learning and plasticity, including the *FoxP2* gene that is essential for speech in humans (Matsunaga and Okanoya, 2009; Wu *et al*., 2021). Furthermore, a comparative analysis of parrot genomes has indicated that several genes related to brain development and cognition in parrots have undergone similar patterns of selection to those observed in human cognition (Wirthlin *et al*., 2018). This is interesting as parrots display remarkable abilities and intelligence to mimic human speech and sound compared to other avian species (Wirthlin *et al*., 2018) .

Whole genome sequences are being made available across all avian orders; certain orders have received extensive genome coverage, including the Passeriformes, Psittaciformes and Galliformes, due to their widespread ecological diversity and/or economic importance (Feng *et al*., 2020; Jarvis *et al*., 2014; Kan *et al*., 2010; Mackiewicz *et al*., 2019; Stiller *et al*., 2024). In contrast, the order of Bucerotiformes, which includes all hornbills and ground hornbills, remains poorly represented in the number of genomes available. To date, the genome sequences of only three hornbill species have been sequenced, namely the Abyssinian Ground Hornbill (*Bucorvus abyssinicus*) (Rhie *et al*., 2021), Great hornbill (*Buceros bicornis*) (Iridian Genomes, 2023), and the Rhinoceros hornbill (*Buceros rhinoceros silvestris*) (Zhang *et al*., 2014). This sparse availability of genome sequences in the order Bucerotiformes compared to other avian orders highlights a significant gap in our understanding of hornbill genetics. The value of genomic data is particularly significant for understanding the evolution of *Bucorvidae* and *Bucerotidae*, given the limited or sparse fossil records and genetic information available for these families (Riamon *et al*., 2021). This genomic information provides a crucial basis for evolutionary studies, compensating for the gaps in the fossil record and offering insights into the genetic and evolutionary history of species in these families.

1.5 Taxonomic classification and characteristics of hornbills

Hornbills belong to the order Bucerotiformes, which is comprised of the families *Bucerotidae* and *Bucorvidae*, with fifteen genera encompassing approximately 61 recognised species (Gonzalez *et al*., 2013; Trail, 2007). Hornbills are known for their large body sizes, vocalisation patterns, self-incarceration of female hornbills during nesting, cooperative breeding patterns and mostly monogamous relationships (Franco and Minggu, 2019; Gonzalez *et al*., 2013; Trail, 2007). Furthermore, they are also well known for their flamboyant bill casques, narrow air-filled cavities enclosed by cancellous bone (bone tissue containing pores) along the upper jaw, which vary in size, shape and coloration (Gamble, 2007). Hornbill species have been placed within their own avian order on the basis of several distinct anatomical features, such as a fused upper vertebra as a result of having a casque, long flattened upper eyelashes and bilobular kidneys (two distinct lobes of the kidney), compared to their sister clades (Feinstein, 1962; Gonzalez *et al*., 2013; Kemp, 2001).

Hornbills are found widely distributed across several regions of Asia and Africa where 31 and twenty-three species are located, respectively. In Asia, 94% of all species occupy tropical rainforests and lead a mainly frugivorous lifestyle, whereas, in Africa, 57% of all hornbill species occupy savannah biomes and thus lead a mainly carnivorous lifestyle (Franco and Minggu, 2019; Viseshakul *et al*., 2011).

The ground hornbills, endemic to Africa, form part of a distinct genus, *Bucorvus* (Kemp, 1995; Vuuren *et al*., 2020). The two species in this genus, *Bucorvus leadbeateri* (Southern Ground Hornbill - SGH) and *Bucorvus abyssinicus* (Abyssinian Ground Hornbill - AGH) (Kemp, 1979) are allopatric to either side of the equator, with the SGH found in the southern regions and the AGH in the northern regions of Africa (Kemp, 1979; Vuuren *et al.*, 2020). The *Bucorvus* species are separated from other hornbills classified in the *Bucerotidae* family as they do not completely seal their females in nest cavities during breeding (Kemp and Kemp, 1980). Both the SGH and AGH species have been classified as Vulnerable by the IUCN (BirdLife International, 2021). As the SGH is the subject of this study, it is discussed in more detail below.

1.6 The Southern Ground Hornbill

1.6.1 Southern Ground Hornbill traits and distribution

The SGH is a large sedentary bird that reaches a body length of 90 to 130 cm and weighs between 2.5 and 6 kg (Kemp, 1995; Kemp, 2005; Theron *et al.*, 2012). It has long legs with short toes and broad wings (Kemp, 1979). Both sexes have black plumages with white primaries (stiff remiges of feathers found on the tip of the wings that are used for flight) and primary coverts (softer feathers covering the main wing feathers for protection), with extensively bare facial features, particularly around the skin of their eyes and inflatable throats (Figure 1.1a) (Hall, 2005; Kemp and Kemp, 1980). This throat skin colouration varies between sexes (Kemp, 1995; Kemp, 2005). Males have red throat skin, while that of females includes a violet-blue patch (Kemp, 1995; Kemp, 2005). Sub-adult SGHs tend to have cream colouration that further develops as they mature (Kemp and Kemp, 1980). The SGH is almost exclusively carnivorous (Kemp, 2005; Kemp and Begg, 1995), lacking the gular pouch found in other hornbill species that allows them to store fruit (Kitamura, 2011). It feeds on a range of invertebrates and small prey which include insects like grasshoppers and beetles, snails, amphibians, lizards, and snakes (Figure 1.1b) (Cilliers *et al.*, 2013; Kemp, 1979). At times they even capture and feed on small mammals and other birds (Carstens *et al.*, 2019).

The SGH is a flagship species which means it is chosen to raise awareness and support for the conservation of biodiversity in its habitat, the African savannah biome (Cilliers *et al.*, 2013; Mabula Ground Hornbill Project, 2019). Other such animal species include vultures, the cheetah and white rhinoceros (Cilliers *et al*., 2013; Mabula Ground Hornbill Project, 2024). The SGH occurs in sixteen African countries, including South Africa, Botswana, Namibia, Mozambique, Zimbabwe, Angola, Zambia, Malawi and Kenya (Daso *et al.*, 2015; Kemp, 1995; Mabula Ground Hornbill Project, 2024). In South Africa, it mainly occurs in the northern and eastern parts of the country in the Limpopo, Mpumalanga, Kwa-Zulu Natal and Eastern Cape provinces but has further been reported in Gauteng and the Free State (Kemp and Webster, 2008). The SGH is highly territorial and occupies year-round territories at densities of about one group per 100 km^2 in South Africa (Kemp *et al.*, 1989).

a)



b)



c)



Figure 1.1: Photographic examples of the SGH. The images represent a) the variation in plumage and throat skin coloration, b) the feeding habits and c) the social structures displayed within the SGH. (Images sourced from the Mabula Ground Hornbill Project, 2024 were taken by a) John Mullineux, b) Anthony Grote Photography and c) Cassie Carsten).

1.6.2 The intricate breeding activity and social structure of the Southern Ground Hornbill

The SGH is one of the largest cooperative breeding birds and they live in groups ranging from two to eleven individuals (Figure 1.1c) (Kemp, 1995; Kemp, 1990). These include the main breeding pair (alpha male and female), juveniles and non-breeding adult members, with the latter typically helper males (Kemp, 1990). SGHs often breed during Austral spring and summer and breeding is usually governed by food availability and rainfall events (Carstens *et al.*, 2019; Kemp, 2005). In the African savannah biome, rainfall events are closely linked to the abundance of invertebrate prey (Kemp, 1976; Kemp and Kemp, 1991). This results in breeding often not occurring annually (Carstens *et al.*, 2019; Kemp and Begg, 1995). Poor conditions lead to the alpha pair forgoing breeding and rather conserving and using their energy to survive and maintain their body conditions, which includes moulting their feathers (Dietz *et al.*, 2013; Langston and Rohwer, 1996). Feathers of SGHs like many other avian species undergo constant mechanical abrasions and degradation as they age, which cannot undergo self-repair (Barta *et al*., 2008; Beltran *et al*., 2018). Such wear and tear affects thermoregulation, aerodynamics and movability (Ausems *et al*., 2021; Kiat, *et al*., 2019). Thus, worn feathers and hairs need to be replaced annually or biannually through the process of moulting (Beltran *et al*., 2018). Moulting itself is a very energy and time costly process and thus conflicts with other energy taxing processes such as breeding and migration (Barta *et al*., 2008).

SGHs are secondary cavity nesters that often nest in natural cavities in trees, rock faces, or earth banks (Kemp and Begg, 1995). These birds also readily accept and nest in artificial nest boxes, which have been found to increase their productivity and survival (Carstens *et al.*, 2019; Wilson and Hockey, 2013). In a study conducted in the Associated Private Nature Reserves (APNR) located next to the Kruger National Park in South Africa, nest box occupancy increased over time between the period 2002 to 2015 (Carstens *et al.*, 2019). Of the 31 nest boxes installed, twenty-three were found to be occupied at least once in the study period. Breeding attempts were also found to be made more often by groups occupying nest boxes than those that occupied natural nest cavities in trees (Carstens *et al.*, 2019). Such nest sites are often reused by SGHs in successive breeding seasons over several years (Kemp and Begg, 1995).

The alpha female typically lays a clutch of one or two (at times three) eggs, three to five days apart (Kemp, 1995). These eggs are then incubated for roughly 42 days by only the alpha female (Kemp, 2005; Kemp and Kemp, 1980). During this time, she is often fed by the alpha

male and helper members of the group (Kemp and Begg, 1995). Despite the number of eggs laid, only one chick is raised and fledged per season even if all the eggs hatch (Kemp, 1995). The second chick in the clutch often dies after a few days post-hatching as it starves to death from parental neglect as well as competition with its older and much larger sibling (Kemp, 1995; Kemp and Kemp, 1980). The chick is fed whole food after it hatches comprising mainly of arthropods (Carstens *et al.*, 2019; Kemp, 1973). The chick fledges after approximately 87 days and leaves the nest (Kemp, 2005; Kemp and Kemp, 1980). Once the chick leaves the nest it is dependent on the group for feeding and is constantly accompanied by a member of the group during the first year for protection (Kemp, 1976; Kemp and Kemp, 1991; Theron *et al.*, 2013).

Juvenile SGHs remain with their parental group for several years, displaying delayed maturity and dependency on the parental group post-fledging (Kemp and Kemp, 1980). Upon maturity, females leave the parental group with non-breeding females floating between home ranges alone or in all-female groups consisting of up to three individuals (Carstens *et al.*, 2019; Kemp, 1990). Males delay leaving the parental group and tend to assist in defending territories and home ranges, detecting predators and rearing future offspring of the alpha pair (Carstens *et al.*, 2019; Kemp, 1990). The first breeding attempt is estimated to occur after these birds reach an age of at least nine years (Kemp, 1995, Kemp, 1976; Kemp and Kemp, 1991). SGHs fledge a chick approximately every two to nine years (Carstens *et al.*, 2019; Kemp, 1990).

Members of this species are generally long-lived, reaching ages of up to 50 years, with the oldest bird in captivity recorded at 70 years of age (Kemp, 1990; Wasser and Sherman, 2010). This is substantially longer than the lifespans of smaller avian species such as the wild house sparrow (*Passer domesticus*) (3 – 4 years in captivity) and the domestic Japanese quail (*Coturnix japonica*) (2 – 3 years) (Harper and Holmes, 2021). However, similar to larger avian species such as the red-crowned crane (*Grus japonensis*), which has an average lifespan of 30 years in the wild and 65 years in captivity, the longevity of the SGH could be correlated to the low metabolic rate requirements of having a larger body size (Lee *et al*., 2020; Speakman, 2005).

1.6.3 Conservation of the Southern Ground Hornbill

SGH populations are declining, with biological factors such as lower reproductive rates and delayed sexual maturity and breeding attempts hampering its ability to recover from threats and re-establish its population (Coetzee *et al*., 2014b). In South Africa, the SGH has lost an estimated 70% of its range and 50% of its total historical populations (Kemp, 2005; Kemp and Webster, 2008). Between 1,290 to 2,380 individuals in southern Africa are estimated to be residing in protected areas of South Africa (Taylor *et al.*, 2015). This includes the Kruger National Park, where 600 to 700 of these individuals are located (Kemp, 2005). The decline in populations has mainly been attributed to the destruction and loss of suitable habitats through anthropogenic activities and bush encroachment (Daso *et al*., 2015; Mabula Ground Hornbill Project, 2024).

Fatalities in SGH populations have also been noted because of indirect poisoning, from the consumption of carcasses laced with poison targeting pests or other carnivores, or from lead from hunting excursions (Koeppel and Kemp, 2015; Mabula Ground Hornbill Project, 2024). In contrast, direct persecution of these birds has been reported, due to their breaking windows as they attack their own reflections, which they perceive as intruding birds due to their territorial behaviour (Kemp and Webster, 2008; Morrison *et al*., 2005). Furthermore, the SGH, similar to many other endangered and protected species such as the rhinoceros, elephants and tigers, have fallen prey to the illicit markets of wildlife and traditional medicine trading (Coetzee *et al*., 2014a; Kemp and Webster, 2008; Morrison *et al*., 2005). Lastly, cultural beliefs, which include the association of the birds with death and evil, predictors of loss, deprivation and destruction, have further resulted in a dwindling in their population numbers (Coetzee *et al*., 2014a, 2014b). In South Africa, projects such as the Mabula Ground Hornbill Project, registered under the BirdLife Preventing Extinctions Programme (BirdLife International), have been working in collaboration with several organisations to enforce conservation efforts and programmes to help save the remaining population of the SGH (Mabula Ground Hornbill Project, 2024). Strategies include educational and awareness campaigns, capturing and rearing of redundant second chicks and reintroducing them to the wild, reintroducing rescued birds in locally extinct and historic ranges, captive breeding programs and providing artificial nests (Jordan, 2011; Mabula Ground Hornbill Project, 2024; Morrison *et al*., 2005). While these conservation strategies have been put into practice for several years, a Population and Habitat Viability Assessment workshop held with over thirty stakeholders, highlighted the lack of

scientific data available regarding these birds (Morrison *et al.*, 2005). As such the need for further research into the biology and ecology of the SGH was recommended (Morrison *et al.*, 2005). This led to the integration of research as another major component of conservation to identify knowledge gaps and strategies that could improve current conservation efforts (Jordan, 2011; Morrison *et al.*, 2005). While SGH research has mainly been focused on the survival, lifestyle, and behavioural patterns of the bird, knowledge regarding its molecular genetics and evolution is still quite limited.

1.7 Rationale of study

To date, genetic research in SGH has primarily centred on resolving the mtDNA control regions and the full mtDNA sequence for phylogenetic analysis with other hornbill species (Delport *et al*., 2002; Gonzalez *et al*., 2013). Phylogenetic analysis of 61 hornbill species using their mtDNA has indicated that the SGH and AGH share a common ancestor with other hornbills. However, these two species are highly divergent members of the order Bucerotiformes, following their own distinct evolutionary path (Gonzalez *et al*., 2013). Additionally, twelve microsatellite markers have been identified and developed (Dalton and Kotze, 2011). More recently, both microsatellite markers and mtDNA have been employed to investigate genetic diversity, population structure and demographic history of the SGH in their habitats across Africa (Kemp *et al*., 2024). This study found that the SGH exhibited low inbreeding, normal gene flow and weak to moderate genetic differentiation amongst the populations studied, largely due to the longevity of these birds and extensive dispersal. These findings have important implications towards the conservation of these species. However, the study also underscored the need for more comprehensive genetic analysis and genomic studies, such as using single nucleotide polymorphisms (SNPs) to develop conservation strategies that can further assist in enhancing the gene flow between populations and ensure the long-term survival of the remaining SGH populations (Kemp *et al*., 2024). To contribute towards these efforts, this study conducted the sequencing, assembly and annotation of the complete genome of the SGH.

1.8 References

Abdul-Muneer, P. M. (2014). Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies. *Genetics Research International*, *2014*, 691759. https://doi.org/10.1155/2014/691759

Ainley, D., Woehler, E., and Lescroel, A. (2016). Birds and Antarctic sea ice. In *Sea Ice* (Third Edition, pp. 570–582). John Wiley and Sons, Ltd. https://doi.org/10.1002/9781118778371.ch24

Akiyama, T., Nishida, C., Momose, K., Onuma, M., Takami, K., and Masuda, R. (2017). Gene duplication and concerted evolution of mitochondrial DNA in crane species. *Molecular Phylogenetics and Evolution*, *106*, 158–163. https://doi.org/10.1016/j.ympev.2016.09.026

Alters, S. (2000). *Biology: Understanding Life* (3rd ed.). Jones and Bartlett Learning.

Anmarkrud, J. A., and Lifjeld, J. T. (2017). Complete mitochondrial genomes of eleven extinct or possibly extinct bird species. *Molecular Ecology Resources*, *17*(2), 334–341. https://doi.org/10.1111/1755-0998.12600

Arcones, A., Ponti, R., and Vieites, D. R. (2021). Mitochondrial substitution rates estimation for divergence time analyses in modern birds based on full mitochondrial genomes. *Ibis*, *163*(4), 1463–1471. https://doi.org/10.1111/ibi.12965

Auer, H., Mayr, B., Lambrou, M., and Schleger, W. (1987). An extended chicken karyotype, including the NOR chromosome. *Cytogenetics and Cell Genetics*, *45*(3–4), 218–221. https://doi.org/10.1159/000132457

Ausems, A. N. M. A., Skrzypek, G., Wojczulanis-Jakubas, K., and Jakubas, D. (2021). Birds of a feather moult together: Differences in moulting distribution of four species of storm-petrels. *PLOS ONE*, *16*(1), e0245756. https://doi.org/10.1371/journal.pone.0245756

B10K. (2024). *B10K Database*. https://b10k.genomics.cn/progress.html

Balanoff, A. M., Bever, G. S., Rowe, T. B., and Norell, M. A. (2013). Evolutionary origins of the avian brain. *Nature*, *501*(7465), Article 7465. https://doi.org/10.1038/nature12424

Baldwin, M. W., Toda, Y., Nakagita, T., O'Connell, M. J., Klasing, K. C., Misaka, T., Edwards, S. V., and Liberles, S. D. (2014). Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. *Science (New York, N.Y.)*, *345*(6199), 929–933. https://doi.org/10.1126/science.1255097

Barcellos, S. A., Kretschmer, R., de Souza, M. S., Costa, A. L., Degrandi, T. M., dos Santos, M. S., de Oliveira, E. H. C., Cioffi, M. B., Gunski, R. J., and Garnero, A. D. V. (2019). Karyotype Evolution and Distinct Evolutionary History of the W Chromosomes in Swallows (Aves, Passeriformes). *Cytogenetic and Genome Research*, *158*(2), 98–105. https://doi.org/10.1159/000500621

Barrett, L. W., Fletcher, S., and Wilton, S. D. (2012). Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements. *Cellular and Molecular Life Sciences*, *69*(21), 3613–3634. https://doi.org/10.1007/s00018-012-0990-9

Barta, Z., McNamara, J. M., Houston, A. I., Weber, T. P., Hedenström, A., and Feró, O. (2008). Optimal moult strategies in migratory birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1490), 211–229. https://doi.org/10.1098/rstb.2007.2136

Bellott, D. W., Skaletsky, H., Pyntikova, T., Mardis, E. R., Graves, T., Kremitzki, C., Brown, L. G., Rozen, S., Warren, W. C., Wilson, R. K., and Page, D. C. (2010). Convergent Evolution of Chicken Z and Human X Chromosomes by Expansion and Gene Acquisition. *Nature*, *466*(7306), 612–616. https://doi.org/10.1038/nature09172

Beltran, R. S., Burns, J. M., and Breed, G. A. (2018). Convergence of biannual moulting strategies across birds and mammals. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1878). https://doi.org/10.1098/rspb.2018.0318

Benton, M., Donoghue, P., Friedman, R., Vinther, T., Asher, R., Friedman, M., and Vinther, J. (2015). Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*, *18*, 1–116. https://doi.org/10.26879/424

Bhullar, B. A. S., Hanson, M., Fabbri, M., Pritchard, A., Bever, G. S., and Hoffman, E. (2016). How to Make a Bird Skull: Major Transitions in the Evolution of the Avian Cranium, Paedomorphosis, and the Beak as a Surrogate Hand. *Integrative and Comparative Biology*, *56*(3), 389–403. https://doi.org/10.1093/icb/icw069

BirdLife International. (2021). *IUCN Red List for birds*. http://www.birdlife.org

BirdLife International. (2022). *State of the World's Birds 2022*. BirdLife International. https://datazone.birdlife.org/userfiles/images/SOWB2022_EN_compressed.pdf

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research*, *27*(8), 1767–1780. https://doi.org/10.1093/nar/27.8.1767

Bravo, G. A., Schmitt, C. J., and Edwards, S. V. (2021). What Have We Learned from the First 500 Avian Genomes? *Annual Review of Ecology, Evolution, and Systematics*, *52*(1), 611–639. https://doi.org/10.1146/annurev-ecolsys-012121-085928

Brusatte, S. L., O'Connor, J. K., and Jarvis, E. D. (2015). The Origin and Diversification of Birds. *Current Biology*, *25*(19), R888–R898. https://doi.org/10.1016/j.cub.2015.08.003

Burt, D. W. (2002). Origin and evolution of avian microchromosomes. *Cytogenetic and Genome Research*, *96*(1–4), 97–112. https://doi.org/10.1159/000063018

Caparroz, R., Rocha, A. V., Cabanne, G. S., Tubaro, P., Aleixo, A., Lemmon, E. M., and Lemmon, A. R. (2018). Mitogenomes of two neotropical bird species and the multiple independent origin of mitochondrial gene orders in Passeriformes. *Molecular Biology Reports*, *45*(3), 279–285. https://doi.org/10.1007/s11033-018-4160-5

Carstens, K. F., Kassanjee, R., Little, R. M., Ryan, P. G., and Hockey, P. A. (2019a). Natal dispersal in the Southern Ground Hornbill *Bucorvus* leadbeateri. *Ostrich*, *90*(2), 119–127. https://doi.org/10.2989/00306525.2019.1590474

Carstens, K. F., Kassanjee, R., Little, R. M., Ryan, P. G., and Hockey, P. a. R. (2019b). Breeding success and population growth of Southern Ground Hornbills *Bucorvus* leadbeateri in an area supplemented with nest-boxes. *Bird Conservation International*, *29*(4), 627–643. https://doi.org/10.1017/S0959270919000108

Carstens, K. F., Kassanjee, R., Little, R. M., Ryan, P. G., and Hockey, P. A. R. (2019c). The effects of weather, group size and type of nest on the timing of egg-laying in the Southern Ground-hornbill *Bucorvus* leadbeateri. *Journal of Ornithology*, *160*(3), 633–640. https://doi.org/10.1007/s10336-019-01654-x

Chen, Y., Xiang, L., Chen, P., and Zhao, H. (2023). Vultures as a model for testing molecular adaptations of dietary specialization in birds. *Avian Research*, *14*, 100128. https://doi.org/10.1016/j.avrs.2023.100128

Cheng, Y., Gao, B., Wang, H., Han, N., Shao, S., Wu, S., Song, G., Zhang, Y. E., Zhu, X., Lu, X., Qu, Y., and Lei, F. (2017). Evolution of beak morphology in the Ground Tit revealed by comparative transcriptomics. *Frontiers in Zoology*, *14*(1), 58. https://doi.org/10.1186/s12983-017-0245-6

Cheng, Y., Miller, M. J., Zhang, D., Song, G., Jia, C., Qu, Y., and Lei, F. (2020). Comparative Genomics Reveals Evolution of a Beak Morphology Locus in a High-Altitude Songbird. *Molecular Biology and Evolution*, *37*(10), 2983–2988. https://doi.org/10.1093/molbev/msaa157

Christidis, L. (1990). *Chrodata 3. B. Aves* (Vol. 4). Animal Cytogenetics.

Cilliers, D., Evans, S., Coetzee, H., and Rensburg, L. van. (2013). Developing a site selection tool to assist reintroduction efforts for the Southern Ground-Hornbill *Bucorvus* leadbeateri. *Ostrich*, *84*(2), 101–111. https://doi.org/10.2989/00306525.2013.823129

Coetzee, H., Nell, W., and Rensburg, L. van. (2014). An Intervention Program Based on Plant Surrogates as Alternatives to the Use of Southern Ground-Hornbills in Cultural Practices. *Ethnobotany Research and Applications*, *12*(0), Article 0.

Coetzee, H., Nell, W., and van Rensburg, L. (2014). An exploration of cultural beliefs and practices across the Southern Ground-Hornbill's range in Africa. *Journal of Ethnobiology and Ethnomedicine*, *10*(1), Article 1. https://doi.org/10.1186/1746-4269-10-28

Dalloul, R. A., Long, J. A., Zimin, A. V., Aslam, L., Beal, K., Blomberg, L. A., Bouffard, P., Burt, D. W., Crasta, O., Crooijmans, R. P. M. A., Cooper, K., Coulombe, R. A., De, S., Delany, M. E., Dodgson, J. B., Dong, J. J., Evans, C., Frederickson, K. M., Flicek, P., …

Reed, K. M. (2010). Multi-Platform Next-Generation Sequencing of the Domestic Turkey (Meleagris gallopavo): Genome Assembly and Analysis. *PLOS Biology*, *8*(9), e1000475. https://doi.org/10.1371/journal.pbio.1000475

Dalton, D., and Kotze, A. (2011). *Isolation and characterization of microsatellite loci in Southern Ground Hornbills (Bucorvus leadbeateri)*.

Damas, J., O'Connor, R. E., Griffin, D. K., and Larkin, D. M. (2019). Avian Chromosomal Evolution. In R. H. S. Kraus (Ed.), *Avian Genomics in Ecology and Evolution: From the Lab into the Wild* (pp. 69–92). Springer International Publishing. https://doi.org/10.1007/978-3-030-16477-5_4

Daso, A. P., Okonkwo, J. O., Jansen, R., Brandao, J. D. D. O., and Kotzé, A. (2015). Mercury concentrations in eggshells of the Southern Ground-Hornbill (*Bucorvus* leadbeateri) and Wattled Crane (Bugeranus carunculatus) in South Africa. *Ecotoxicology and Environmental Safety*, *114*, 61–66. https://doi.org/10.1016/j.ecoenv.2015.01.010

de Koning, A. P. J., Gu, W., Castoe, T. A., Batzer, M. A., and Pollock, D. D. (2011). Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics*, *7*(12), e1002384. https://doi.org/10.1371/journal.pgen.1002384

de Oliveira Furo, I., Kretschmer, R., dos Santos, M. S., de Lima Carvalho, C. A., Gunski, R. J., O'Brien, P. C. M., Ferguson-Smith, M. A., Cioffi, M. B., and de Oliveira, E. H. C. (2017). Chromosomal Mapping of Repetitive DNAs in Myiopsitta monachus and Amazona aestiva (Psittaciformes, Psittacidae) with Emphasis on the Sex Chromosomes. *Cytogenetic and Genome Research*, *151*(3), 151–160. https://doi.org/10.1159/000464458

de Souza, M. S., Kretschmer, R., Barcellos, S. A., Costa, A. L., Cioffi, M. de B., de Oliveira, E. H. C., Garnero, A. D. V., and Gunski, R. J. (2020). Repeat Sequence Mapping Shows Different W Chromosome Evolutionary Pathways in Two Caprimulgiformes Families. *Birds*, *1*(1), Article 1. https://doi.org/10.3390/birds1010004

Dean, W., and Williams, J. (2004). Adaptations of birds for life in deserts with particular reference to Larks (ALAUDIDAE). *Transactions of the Royal Society of South Africa*, *59*, 79–91. https://doi.org/10.1080/00359190409519166

Delbridge, M. L., and Graves, J. A. (1999). Mammalian Y chromosome evolution and the male-specific functions of Y chromosome-borne genes. *Reviews of Reproduction*, *4*(2), 101–109. https://doi.org/10.1530/ror.0.0040101

Delport, W., Ferguson, J. W. H., and Bloomer, P. (2002). Characterization and Evolution of the Mitochondrial DNA Control Region in Hornbills (Bucerotiformes). *Journal of Molecular Evolution*, *54*(6), 794–806. https://doi.org/10.1007/s00239-001-0083-0

Desjardins, P., and Morais, R. (1990). Sequence and gene organization of the chicken mitochondrial genome: A novel gene order in higher vertebrates. *Journal of Molecular Biology*, *212*(4), 599–634. https://doi.org/10.1016/0022-2836(90)90225-B

Dietz, M. W., Rogers, K. G., and Piersma, T. (2013). When the seasons don't fit: Speedy molt as a routine carry-over cost of reproduction. *PloS One*, *8*(1), e53890. https://doi.org/10.1371/journal.pone.0053890

Dodson, P. (2000). Origin of Birds: The Final Solution?1. *American Zoologist*, *40*(4), 504–512. https://doi.org/10.1093/icb/40.4.504

Donsker, G. F. D., and Rasmussen, P. (2023). *IOC World Bird List (v13.2.): Vol. 14.1* (Eds). doi : 10.14344/IOC.ML.13.1.

Feinstein, B. (1962). Additional Cases of Bilobated Kidneys in the Hornbills. *The Auk*, *79*(4), 709–711. https://doi.org/10.2307/4082654

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., … Zhang, G. (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature*, *587*(7833), Article 7833. https://doi.org/10.1038/s41586-020-2873-9

Franco, F. M., and Minggu, M. J. (2019). When the seeds sprout, the hornbills hatch: Understanding the traditional ecological knowledge of the Ibans of Brunei Darussalam on hornbills. *Journal of Ethnobiology and Ethnomedicine*, *15*, 46. https://doi.org/10.1186/s13002-019-0325-0

Fridolfsson, A. K., Cheng, H., Copeland, N. G., Jenkins, N. A., Liu, H. C., Raudsepp, T., Woodage, T., Chowdhary, B., Halverson, J., and Ellegren, H. (1998). Evolution of the avian sex chromosomes from an ancestral pair of autosomes. *Proceedings of the National Academy of Sciences*, *95*(14), 8147–8152. https://doi.org/10.1073/pnas.95.14.8147

Gamble, K. C. (2007). Internal Anatomy of the Hornbill Casque Described by Radiography, Contrast Radiography, and Computed Tomography. *Journal of Avian Medicine and Surgery*, *21*(1), 38–49. https://doi.org/10.1647/1082-6742(2007)21[38:IAOTHC]2.0.CO;2

Gatesy, S. M., and Baier, D. B. (2005). The origin of the avian flight stroke: A kinematic and kinetic perspective. *Paleobiology*, *31*(3), 382–399. https://doi.org/10.1666/0094-8373(2005)031[0382:TOOTAF]2.0.CO;2

Gatesy, S. M., and Dial, K. P. (1996). From Frond to Fan: Archaeopteryx and the Evolution of Short-Tailed Birds. *Evolution*, *50*(5), 2037–2048. https://doi.org/10.1111/j.1558-5646.1996.tb03590.x

Gemmell, N. J., Rutherford, K., Prost, S., Tollis, M., Winter, D., Macey, J. R., Adelson, D. L., Suh, A., Bertozzi, T., Grau, J. H., Organ, C., Gardner, P. P., Muffato, M., Patricio, M., Billis, K., Martin, F. J., Flicek, P., Petersen, B., Kang, L., … Stone, C. (2020). The tuatara genome reveals ancient features of amniote evolution. *Nature*, *584*(7821), 403–409. https://doi.org/10.1038/s41586-020-2561-9

Gibb, G. C., Kardailsky, O., Kimball, R. T., Braun, E. L., and Penny, D. (2007). Mitochondrial Genomes and Avian Phylogeny: Complex Characters and Resolvability without Explosive Radiations. *Molecular Biology and Evolution*, *24*(1), 269–280. https://doi.org/10.1093/molbev/msl158

Gonzalez, J. C. T., Sheldon, B. C., Collar, N. J., and Tobias, J. A. (2013). A comprehensive molecular phylogeny for the hornbills (Aves: Bucerotidae). *Molecular Phylogenetics and Evolution*, *67*(2), 468–483. https://doi.org/10.1016/j.ympev.2013.02.012

Grealy, A., Bunce, M., and Holleley, C. E. (2019). Avian mitochondrial genomes retrieved from museum eggshell. *Molecular Ecology Resources*, *19*(4), 1052–1062. https://doi.org/10.1111/1755-0998.13007

Gregory, T. R. (2001). Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biological Reviews*, *76*(1), 65–101. https://doi.org/10.1017/S1464793100005595

Gregory, T. R., Andrews, C. B., McGuire, J. A., and Witt, C. C. (2009). The smallest avian genomes are found in hummingbirds. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1674), 3753–3757. https://doi.org/10.1098/rspb.2009.1004

Gregory, T. R., Nicol, J. A., Tamm, H., Kullman, B., Kullman, K., Leitch, I. J., Murray, B. G., Kapraun, D. F., Greilhuber, J., and Bennett, M. D. (2007). Eukaryotic genome size databases. *Nucleic Acids Research*, *35*(Database issue), D332–D338. https://doi.org/10.1093/nar/gkl828

Gunski, R. J., Kretschmer, R., Santos de Souza, M., de Oliveira Furo, I., Barcellos, S. A., Costa, A. L., Cioffi, M. B., de Oliveira, E. H. C., and del Valle Garnero, A. (2019). Evolution of Bird Sex Chromosomes Narrated by Repetitive Sequences: Unusual W Chromosome Enlargement in Gallinula melanops (Aves: Gruiformes: Rallidae). *Cytogenetic and Genome Research*, *158*(3), 152–159. https://doi.org/10.1159/000501381

Hall, K. S. S. (2005). Do nine-primaried passerines have nine or ten primary feathers? The evolution of a concept. *Journal of Ornithology*, *146*(2), 121–126. https://doi.org/10.1007/s10336-004-0070-5

Handley, L. J. L., Ceplitis, H., and Ellegren, H. (2004). Evolutionary Strata on the Chicken Z Chromosome: Implications for Sex Chromosome Evolution. *Genetics*, *167*(1), 367–376. https://doi.org/10.1534/genetics.167.1.367

Harper, J. M., and Holmes, D. J. (2021). New Perspectives on Avian Models for Studies of Basic Aging Processes. *Biomedicines*, *9*(6), Article 6. https://doi.org/10.3390/biomedicines9060649

Harrison, G. L. (Abby), McLenachan, P. A., Phillips, M. J., Slack, K. E., Cooper, A., and Penny, D. (2004). Four New Avian Mitochondrial Genomes Help Get to Basic Evolutionary

Questions in the Late Cretaceous. *Molecular Biology and Evolution*, *21*(6), 974–983. https://doi.org/10.1093/molbev/msh065

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., and Guigó, R. (2009). Identifying protein-coding genes in genomic sequences. *Genome Biology*, *10*(1), 201. https://doi.org/10.1186/gb-2009-10-1-201

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A. M., Delany, M. E., Dodgson, J. B., Chinwalla, A. T., Cliften, P. F., Clifton, S. W., Delehaunty, K. D., Fronick, C., Fulton, R. S., Graves, T. A., Kremitzki, C., … Project management: (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), Article 7018. https://doi.org/10.1038/nature03154

Huang, Z., De O. Furo, I., Liu, J., Peona, V., Gomes, A. J. B., Cen, W., Huang, H., Zhang, Y., Chen, D., Xue, T., Zhang, Q., Yue, Z., Wang, Q., Yu, L., Chen, Y., Suh, A., de Oliveira, E. H. C., and Xu, L. (2022). Recurrent chromosome reshuffling and the evolution of neo-sex chromosomes in parrots. *Nature Communications*, *13*, 944. https://doi.org/10.1038/s41467-022-28585-1

Hughes, A. L. (1999). *Adaptive Evolution of Genes and Genomes*. Oxford University Press.

Hughes, A. L., and Hughes, M. K. (1995). Small genomes for better flyers. *Nature*, *377*(6548), Article 6548. https://doi.org/10.1038/377391a0

Hughes, A. L., and Piontkivska, H. (2005). DNA repeat arrays in chicken and human genomes and the adaptive evolution of avian genome size. *BMC Evolutionary Biology*, *5*(1), 12. https://doi.org/10.1186/1471-2148-5-12

Hutchinson, J. R. (2001). The evolution of pelvic osteology and soft tissues on the line to extant birds (Neornithes). *Zoological Journal of the Linnean Society*, *131*(2), 123–168. https://doi.org/10.1111/j.1096-3642.2001.tb01313.x

Iridian Genomes. (2023). *IRIDIAN GENOMES*. https://www.iridiangenomes.com/contact.html

Ishijima, J., Uno, Y., Nishida, C., and Matsuda, Y. (2014). Genomic Structures of the kW1 Loci on the Z and W Chromosomes in Ratite Birds: Structural Changes at an Early Stage of

W Chromosome Differentiation. *Cytogenetic and Genome Research*, *142*(4), 255–267. https://doi.org/10.1159/000362479

Janitz, K., and Janitz, M. (2011). Chapter 12—Assessing Epigenetic Information. In T. Tollefsbol (Ed.), *Handbook of Epigenetics* (pp. 173–181). Academic Press. https://doi.org/10.1016/B978-0-12-375709-8.00012-5

Janke, A., and Arnason, U. (1997). The complete mitochondrial genome of Alligator mississippiensis and the separation between recent archosauria (birds and crocodiles). *Molecular Biology and Evolution*, *14*(12), 1266–1272. https://doi.org/10.1093/oxfordjournals.molbev.a025736

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Alfaro-Núñez, A., Narula, N., Liu, L., Burt, D., Ellegren, H., Edwards, S. V., … The Avian Phylogenomics Consortium. (2015). Phylogenomic analyses data of the avian phylogenomics project. *GigaScience*, *4*(1), s13742-014-0038–1. https://doi.org/10.1186/s13742-014-0038-1

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., … Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science (New York, N.Y.)*, *346*(6215), 1320–1331. https://doi.org/10.1126/science.1253451

Jax, E., Wink, M., and Kraus, R. H. S. (2018). Avian transcriptomics: Opportunities and challenges. *Journal of Ornithology*, *159*(3), 599–629. https://doi.org/10.1007/s10336-018-1532-5

Jordan, M. (2011). *Southern Ground Hornbill (Bucorvus leadbeateri) Species Recovery Plan for South Africa*. Johannesburg Zoo/Endangered Wildlife Trust. https://doi.org/10.13140/RG.2.1.3554.0082

Kan, X. Z., Yang, J. K., Li, X. F., Chen, L., Lei, Z. P., Wang, M., Qian, C. J., Gao, H., and Yang, Z. Y. (2010). Phylogeny of major lineages of galliform birds (Aves: Galliformes) based on complete mitochondrial genomes. *Genetics and Molecular Research: GMR*, *9*(3), 1625–1633. https://doi.org/10.4238/vol9-3gmr898

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), 164–185. https://doi.org/10.1111/nyas.13295

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, *114*(8), E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Kemp, A. (1976). Factors affecting the onset of breeding in African hornbills. *Proceedings of the 16 Th International Ornithological Congress*, 248–257.

Kemp, A. (1995). *The Hornbills: Bucerotiformes* (Vol. 114). Oxford University Press.

Kemp, A. (2005). Southern Ground Hornbill. In *In Hockey, P. A. R., Dean, W. R. J., and Ryan, P. G. (eds). Roberts Birds of Southern Africa* (7th ed., pp. 158–159). Trustees of the John Voelcker Bird Book Fund.

Kemp, A. C. (1973). Environmental factors affecting the onset of breeding in some southern African hornbills, Tockus spp. *Journal of Reproduction and Fertility. Supplement*, *19*, 319–331.

Kemp, A. C. (1979). A review of the hornbills: Biology and radiation. In *The living bird* (Seventeenth Annual, pp. 105–136). Cornell Laboratory of Ornithology. https://docs.google.com/file/d/0B5qf0llRHM7PalAwLUxfaVJEMU0/edit

Kemp, A. C. (1990). The behavioural ecology of the Southern Ground Hornbill: Are competitive offspring at a premium? *Proceedings of International 100 Deutscheornitologen-Gesselschaft Meeting Current Topics in Avian Biology*, *1988*, 267–271.

Kemp, A. C. (2001). Family Bucerotidae (Hornbills). In *Del Hoyo, J., Elliott, A. Sargatal, J. (Eds.), Handbook of the Birds of the World, Piciformes to Bucerotiformes* (Vol. 6, pp. 436–523). Lynx Edicions.

Kemp, A. C., and Begg, K. S. (1995). Nest Sites of the Southern Ground Hornbill *Bucorvus* Leadbeateri in the Kruger National Park, South Africa, and Conservation Implications. *Ostrich*, *67*(1), 9–14. https://doi.org/10.1080/00306525.1996.9633773

Kemp, A. C., Joubert, S. C. J., and Kemp, M. I. (1989). Distribution of southern ground horn bills in the Kruger National Park in relation to some environmental features. *South African Journal of Wildlife Research - 24-Month Delayed Open Access*, *19*(3), 93–98. https://doi.org/10.10520/AJA03794369_1688

Kemp, A. C., and Kemp, M. I. (1980). The biology of the Southern Ground Hornbill *Bucorvus* leadbeateri (Vigors) (Aves: Bucerotidae). *Annals of the Transvaal Museum*, *32*(4), 65–100. https://doi.org/10.10520/AJA00411752_1057

Kemp, A. C., and Kemp, M. I. (1991). Timing of egg laying by Southern Ground-Hornbills *Bucorvus* leadbeateri in the central Kruger National Park, South Africa. *Ostrich*, *62*, 80–82.

Kemp, A., and Webster, R. (2008). *Latest analysis of Southern Ground Hornbill (SGH) distribution and population in South Africa: December 2008* (pp. 1–11) [Unpublished report]. https://docs.google.com/file/d/0B5qf0llRHM7PT2I1bUNDaTRQcXM/edit

Kemp, L., Dalton, D. L., Mwale, M., Paul Grobler, J., Thabang Madisha, M., van Wyk, A. M., Mokgokong, P. S., Jansen, R., and Kotze, A. (2024). Broad-scale genetic assessment of Southern Ground-Hornbills (*Bucorvus leadbeateri)* to inform population management. *Global Ecology and Conservation*, *52*, e02963. https://doi.org/10.1016/j.gecco.2024.e02963

Kiat, Y., Vortman, Y., and Sapir, N. (2019). Feather moult and bird appearance are correlated with global warming over the last 200 years. *Nature Communications*, *10*(1), Article 1. https://doi.org/10.1038/s41467-019-10452-1

Kitamura, S. (2011). Frugivory and seed dispersal by hornbills (Bucerotidae) in tropical forests. *Acta Oecologica*, *37*(6), 531–541. https://doi.org/10.1016/j.actao.2011.01.015

Koepfli, K. P., Paten, B., and and O'Brien, S. J. (2015). The Genome 10K Project: A Way Forward. *Annual Review of Animal Biosciences*, *3*(1), 57–111. https://doi.org/10.1146/annurev-animal-090414-014900

Koeppel, K. N., and Kemp, L. V. (2015). Lead Toxicosis in a Southern Ground Hornbill *Bucorvus* leadbeateri in South Africa. *Journal of Avian Medicine and Surgery*, *29*(4), 340–344. https://doi.org/10.1647/2014-037

Kordis, D. (2010). Transposable Elements in Reptilian and Avian (Sauropsida) Genomes. *Cytogenetic and Genome Research*, *127*(2–4), 94–111. https://doi.org/10.1159/000294999

Krishan, A. (1964). Microchromosomes in the spermatogenesis of the domestic turkey. *Experimental Cell Research*, *33*(1), 1–7. https://doi.org/10.1016/S0014-4827(64)81004-1

Lamichhaney, S., Berglund, J., Almén, M. S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerová, M., Rubin, C. J., Wang, C., Zamani, N., Grant, B. R., Grant, P. R., Webster, M. T., and Andersson, L. (2015). Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nature*, *518*(7539), 371–375. https://doi.org/10.1038/nature14181

Lamichhaney, S., Han, F., Berglund, J., Wang, C., Almén, M. S., Webster, M. T., Grant, B. R., Grant, P. R., and Andersson, L. (2016). A beak size locus in Darwin's finches facilitated character displacement during a drought. *Science*, *352*(6284), 470–474. https://doi.org/10.1126/science.aad8786

Langston, N. E., and Rohwer, S. (1996). Molt-Breeding Tradeoffs in Albatrosses: Life History Implications for Big Birds. *Oikos*, *76*(3), 498–510. https://doi.org/10.2307/3546343

Lee, H., Kim, J., Weber, J. A., Chung, O., Cho, Y. S., Jho, S., Jun, J., Kim, H. M., Lim, J., Choi, J. P., Jeon, S., Blazyte, A., Edwards, J. S., Paek, W. K., and Bhak, J. (2020). Whole Genome Analysis of the Red-Crowned Crane Provides Insight into Avian Longevity. *Molecules and Cells*, *43*(1), 86–95. https://doi.org/10.14348/molcells.2019.0190

Liang, H., McConkey, K. R., and Lim, J. Y. (2024). From beak to fruit: An Asian hornbill database for frugivory and seed dispersal research. *Global Ecology and Conservation*, *51*, e02879. https://doi.org/10.1016/j.gecco.2024.e02879

Liu, Y., Li, G., Guo, Z., Zhang, H., Wei, B., and He, D. (2024). Transcriptome analysis of sexual dimorphism in dorsal down coloration in goslings. *BMC Genomics*, *25*(1), 505. https://doi.org/10.1186/s12864-024-10394-z

Liu, J., Wang, Z., Li, J., Xu, L., Liu, J., Feng, S., Guo, C., Chen, S., Ren, Z., Rao, J., Wei, K., Chen, Y., Jarvis, E. D., Zhang, G., and Zhou, Q. (2021). A new emu genome illuminates the evolution of genome configuration and nuclear architecture of avian chromosomes. *Genome Research*, *31*(3), 497–511. https://doi.org/10.1101/gr.271569.120

Lovell, P. V., Wirthlin, M., Wilhelm, L., Minx, P., Lazar, N. H., Carbone, L., Warren, W. C., and Mello, C. V. (2014). Conserved syntenic clusters of protein coding genes are missing in birds. *Genome Biology*, *15*(12), 565. https://doi.org/10.1186/s13059-014-0565-1

Mabula Ground-Hornbill Project. (2024). *Mabula Ground-Hornbill Project—All you need to know*. https://ground-hornbill.org.za/about-ground-hornbills

Mackiewicz, P., Urantówka, A. D., Kroczak, A., and Mackiewicz, D. (2019). Resolving Phylogenetic Relationships within Passeriformes Based on Mitochondrial Genes and Inferring the Evolution of Their Mitogenomes in Terms of Duplications. *Genome Biology and Evolution*, *11*(10), 2824–2849. https://doi.org/10.1093/gbe/evz209

Mank, J. E., and Ellegren, H. (2007). Parallel divergence and degradation of the avian W sex chromosome. *Trends in Ecology and Evolution*, *22*(8), 389–391. https://doi.org/10.1016/j.tree.2007.05.003

Marshall, A. J. (2013). *Biology and Comparative Physiology of Birds: Volume I* (Vol. 1). Academic Press.

Marugán-Lobón, J., Watanabe, A., and Kawabe, S. (2016). Studying avian encephalization with geometric morphometrics. *Journal of Anatomy*, *229*(2), 191–203. https://doi.org/10.1111/joa.12476

Matsunaga, E., and Okanoya, K. (2009). Evolution and diversity in avian vocal system: An Evo-Devo model from the morphological and behavioral perspectives. *Development, Growth and Differentiation*, *51*(3), 355–367. https://doi.org/10.1111/j.1440-169X.2009.01091.x

McQueen, H. A., Siriaco, G., and Bird, A. P. (1998). Chicken Microchromosomes Are Hyperacetylated, Early Replicating, and Gene Rich. *Genome Research*, *8*(6), 621–630.

Mehrotra, S., and Goyal, V. (2014). Repetitive Sequences in Plant Nuclear DNA: Types, Distribution, Evolution and Function. *Genomics, Proteomics and Bioinformatics*, *12*(4), 164–171. https://doi.org/10.1016/j.gpb.2014.07.003

Mindell, D. P., Sorenson, M. D., and Dimcheff, D. E. (1998). Multiple independent origins of mitochondrial gene order in birds. *Proceedings of the National Academy of Sciences*, *95*(18), 10693–10697. https://doi.org/10.1073/pnas.95.18.10693

Morrison, K., Daly, B., Burden, D., Engelbrecht, D., Jordan, M., Kemp, M., Potgieter, C., Turner, A., and Friedmann, Y., (eds). (2005). *Southern Ground Hornbill (Bucorvus leadbeateri) PHVA Workshop Report*. Conservation Breeding Specialist Group (SSC / IUCN) / CBSG South Africa; Endangered Wildlife Trust. https://doi.org/10.13140/RG.2.1.4752.9442

O'Connor, J. (2022). Enantiornithes. *Current Biology*, *32*(20), R1166–R1172. https://doi.org/10.1016/j.cub.2022.07.074

Ohta, T. (1992). The Nearly Neutral Theory of Molecular Evolution. *Annual Review of Ecology and Systematics*, *23*, 263–286.

Oliveira, T. D. de, Kretschmer, R., Bertocchi, N. A., Degrandi, T. M., Oliveira, E. H. C. de, Cioffi, M. de B., Garnero, A. del V., and Gunski, R. J. (2017). Genomic Organization of Repetitive DNA in Woodpeckers (Aves, Piciformes): Implications for Karyotype and ZW Sex Chromosome Differentiation. *PLOS ONE*, *12*(1), e0169987. https://doi.org/10.1371/journal.pone.0169987

Padian, K., and Chiappe, L. M. (1998). The Origin of Birds and Their Flight. *Scientific American*, *278*(2), 38–47.

Rafferty, A. R., and Reina, R. D. (2012). Arrested embryonic development: A review of strategies to delay hatching in egg-laying reptiles. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1737), 2299–2308. https://doi.org/10.1098/rspb.2012.0100

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D.

(2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), Article 7856. https://doi.org/10.1038/s41586-021-03451-0

Riamon, S., Pickford, M., Senut, B. and Louchart, A. (2021). Bucerotidae from the early Miocene of Napak, Uganda (East Africa): the earliest hornbill with a modern-type beak. *Ibis*, *163*(2), 715–721. https://doi.org/10.1111/ibi.12907

Saha, A., Bellucci, A., Fratini, S., Cannicci, S., Ciofi, C., and Iannucci, A. (2023). Ecological factors and parity mode correlate with genome size variation in squamate reptiles. *BMC Ecology and Evolution*, *23*(1), 69. https://doi.org/10.1186/s12862-023-02180-4

Sammler, S., Bleidorn, C., and Tiedemann, R. (2011). Full mitochondrial genome sequences of two endemic Philippine hornbill species (Aves: Bucerotidae) provide evidence for pervasive mitochondrial DNA recombination. *BMC Genomics*, *12*, 35. https://doi.org/10.1186/1471-2164-12-35

Scanes, C. G. (2020). Avian Physiology: Are Birds Simply Feathered Mammals? *Frontiers in Physiology*, *11*, 542466. https://doi.org/10.3389/fphys.2020.542466

Sfetcu, N. (2014). *The Birds World* (1st ed.). MultiMedia Publishing.

Shabalina, S. A., and Spiridonov, N. A. (2004). The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, *5*(4), 105.

Simna, S. P., and Han, Z. (2022). Prospects of Non-Coding Elements in Genomic DNA Based Gene Therapy. *Current Gene Therapy*, *22*(2), 89. https://doi.org/10.2174/1566523221666210419090357

Slack, K. E., Delsuc, F., Mclenachan, P. A., Arnason, U., and Penny, D. (2007). Resolving the root of the avian mitogenomic tree by breaking up long branches. *Molecular Phylogenetics and Evolution*, *42*(1), 1–13. https://doi.org/10.1016/j.ympev.2006.06.002

Slack, K. E., Jones, C. M., Ando, T., Harrison, G. L. (Abby), Fordyce, R. E., Arnason, U., and Penny, D. (2006). Early Penguin Fossils, Plus Mitochondrial Genomes, Calibrate Avian

Evolution. *Molecular Biology and Evolution*, *23*(6), 1144–1155. https://doi.org/10.1093/molbev/msj124

Smeds, L., Kawakami, T., Burri, R., Bolivar, P., Husby, A., Qvarnström, A., Uebbing, S., and Ellegren, H. (2014). Genomic identification and characterization of the pseudoautosomal region in highly differentiated avian sex chromosomes. *Nature Communications*, *5*, 5448. https://doi.org/10.1038/ncomms6448

Smith, C. A., Roeszler, K. N., Ohnesorg, T., Cummins, D. M., Farlie, P. G., Doran, T. J., and Sinclair, A. H. (2009). The avian Z-linked gene DMRT1 is required for male sex determination in the chicken. *Nature*, *461*(7261), 267–271. https://doi.org/10.1038/nature08298

Smith, J., Bruley, C. K., Paton, I. R., Dunn, I., Jones, C. T., Windsor, D., Morrice, D. R., Law, A. S., Masabanda, J., Sazanov, A., Waddington, D., Fries, R., and Burt, D. W. (2000). Differences in gene density on chicken macrochromosomes and microchromosomes. *Animal Genetics*, *31*(2), 96–103. https://doi.org/10.1046/j.1365-2052.2000.00565.x

Solovei, I., Gaginskaya, E. R., and Macgregor, H. C. (1994). The arrangement and transcription of telomere DNA sequences at the ends of lampbrush chromosomes of birds. *Chromosome Research: An International Journal on the Molecular, Supramolecular and Evolutionary Aspects of Chromosome Biology*, *2*(6), 460–470. https://doi.org/10.1007/BF01552869

Speakman, J. R. (2005). Body size, energy metabolism and lifespan. *Journal of Experimental Biology*, *208*(9), 1717–1730. https://doi.org/10.1242/jeb.01556

Sterner, D. E., and Berger, S. L. (2000). Acetylation of Histones and Transcription-Related Factors. *Microbiology and Molecular Biology Reviews*, *64*(2), 435–459.

Stiglec, R., Ezaz, T., and Graves, J. A. M. (2007). A new look at the evolution of avian sex chromosomes. *Cytogenetic and Genome Research*, *117*(1–4), 103–109. https://doi.org/10.1159/000103170

Stiller, J., Feng, S., Chowdhury, A. A., Rivas-González, I., Duchêne, D. A., Fang, Q., Deng, Y., Kozlov, A., Stamatakis, A., Claramunt, S., Nguyen, J. M. T., Ho, S. Y. W., Faircloth, B. C., Haag, J., Houde, P., Cracraft, J., Balaban, M., Mai, U., Chen, G., … Zhang, G.

(2024). Complexity of avian evolution revealed by family-level genomes. *Nature*, 1–10. https://doi.org/10.1038/s41586-024-07323-1

Stiller, J., and Zhang, G. (2019). Comparative Phylogenomics, a Stepping Stone for Bird Biodiversity Studies. *Diversity*. https://agris.fao.org/agris-search/search.do?recordID=US201900404175

Szarski, H. (1976). Cell Size and Nuclear DNA Content in Vertebrates. In G. H. Bourne, J. F. Danielli, and K. W. Jeon (Eds.), *International Review of Cytology* (Vol. 44, pp. 93–111). Academic Press. https://doi.org/10.1016/S0074-7696(08)61648-4

Takagi, N., and Sasaki, M. (1974). A phylogenetic study of bird karyotypes. *Chromosoma*, *46*(1), 91–120. https://doi.org/10.1007/BF00332341

Taylor, M. R., Peacock, F., and Wanless, R. M. (eds). (2015). Southern Ground Hornbill. In *The 2015 Eskom Red Data Book of Birds of South Africa, Lesotho and Swaziland* (pp. 119–121). BirdLife South Africa.

Theron, N., Dalton, D., Grobler, J. P., Jansen, R., and Kotze, A. (2013). Molecular insights on the re-colonization of the Limpopo Valley, South Africa, by Southern Ground-Hornbills. *Journal of Ornithology*, *154*(3), 727–737. https://doi.org/10.1007/s10336-013-0937-4

Theron, N., Jansen, R., Grobler, J., and Kotze, A. (2012). The home range of a recently established group of Southern ground-hornbill (*Bucorvus* leadbeateri) in the Limpopo Valley, South Africa. *Koedoe*, *55*, 1–8. https://doi.org/10.4102/koedoe.v55i1.1135

Torgasheva, A., Malinovskaya, L., Zadesenets, K. S., Slobodchikova, A., Shnaider, E., Rubtsov, N., and Borodin, P. (2021). Highly Conservative Pattern of Sex Chromosome Synapsis and Recombination in Neognathae Birds. *Genes*, *12*(9), Article 9. https://doi.org/10.3390/genes12091358

Trail, P. (2007). African hornbills: Keystone species threatened by habitat loss, hunting and international trade. *Ostrich - Journal of African Ornithology*, *October 2007*, 609–613. https://doi.org/10.2989/OSTRICH.2007.78.3.7.318

Verkuil, Y. I., Piersma, T., and Baker, A. J. (2010). A novel mitochondrial gene order in shorebirds (Scolopacidae, Charadriiformes). *Molecular Phylogenetics and Evolution*, *57*(1), 411–416. https://doi.org/10.1016/j.ympev.2010.06.010

Viseshakul, N., Charoennitikul, W., Kitamura, S., Kemp, A., Thong-Aree, S., Surapunpitak, Y., Poonswad, P., and Ponglikitmongkol, M. (2011). A phylogeny of frugivorous hornbills linked to the evolution of Indian plants within Asian rainforests. *Journal of Evolutionary Biology*, *24*(7), 1533–1545. https://doi.org/10.1111/j.1420-9101.2011.02285.x

Vuuren, A. K. J. van, Kemp, L. V., and McKechnie, A. E. (2020). The beak and unfeathered skin as heat radiators in the southern ground-hornbill. *Journal of Avian Biology*, *51*(6), Article 6. https://doi.org/10.1111/jav.02457

Warren, W. C., Clayton, D. F., Ellegren, H., Arnold, A. P., Hillier, L. W., Künstner, A., Searle, S., White, S., Vilella, A. J., Fairley, S., Heger, A., Kong, L., Ponting, C. P., Jarvis, E. D., Mello, C. V., Minx, P., Lovell, P., Velho, T. A. F., Ferris, M., … Wilson, R. K. (2010). The genome of a songbird. *Nature*, *464*(7289), 757–762. https://doi.org/10.1038/nature08819

Wasser, D. E., and Sherman, P. W. (2010). Avian longevities and their interpretation under evolutionary theories of senescence. *Journal of Zoology*, *280*(2), 103–155. https://doi.org/10.1111/j.1469-7998.2009.00671.x

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., Georges, A., and Graves, J. A. M. (2021). Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proceedings of the National Academy of Sciences*, *118*(45), e2112494118. https://doi.org/10.1073/pnas.2112494118

Whelan, C. J., Wenny, D. G., and Marquis, R. J. (2008). Ecosystem services provided by birds. *Annals of the New York Academy of Sciences*, *1134*, 25–60. https://doi.org/10.1196/annals.1439.003

Whitehouse, A. (2018). *Evolution Unredacted*. Bilquees Press.

Wilson, G., and Hockey, P. A. R. (2013). Causes of variable reproductive performance by Southern Ground-hornbill *Bucorvus* leadbeateri and implications for management. *Ibis*, *155*(3), 476–484. https://doi.org/10.1111/ibi.12042

Wirthlin, M., Lima, N. C. B., Guedes, R. L. M., Soares, A. E. R., Almeida, L. G. P., Cavaleiro, N. P., Loss de Morais, G., Chaves, A. V., Howard, J. T., Teixeira, M. de M., Schneider, P. N., Santos, F. R., Schatz, M. C., Felipe, M. S., Miyaki, C. Y., Aleixo, A., Schneider, M. P. C., Jarvis, E. D., Vasconcelos, A. T. R., … Mello, C. V. (2018). Parrot Genomes and the Evolution of Heightened Longevity and Cognition. *Current Biology*, *28*(24), 4001-4008.e7. https://doi.org/10.1016/j.cub.2018.10.050

Wu, L., Jiao, X., Zhang, D., Cheng, Y., Song, G., Qu, Y., and Lei, F. (2021). Comparative Genomics and Evolution of Avian Specialized Traits. *Current Genomics*, *22*(7), 496–511. https://doi.org/10.2174/1389202923666211227143952

Wu, Y., and Wang, H. (2019). Convergent evolution of bird-mammal shared characteristics for adapting to nocturnality. *Proceedings of the Royal Society B: Biological Sciences*, *286*(1897), 20182185. https://doi.org/10.1098/rspb.2018.2185

Yan, Q., Korty, R., Zhang, Z., Brierley, C., Li, X., and Wang, H. (2020). Large shift of the Pacific Walker Circulation across the Cenozoic. *National Science Review*, *8*(5), nwaa101. https://doi.org/10.1093/nsr/nwaa101

Yuan, Q., Guo, Q., Cao, J., Luo, X., and Duan, Y. (2023). Description of the Three Complete Mitochondrial Genomes of Sitta (S. himalayensis, S. nagaensis, and S. yunnanensis) and Phylogenetic Relationship (Aves: Sittidae). *Genes*, *14*(3), 589. https://doi.org/10.3390/genes14030589

Zepeda Mendoza, M. L., Roggenbuck, M., Manzano Vargas, K., Hansen, L. H., Brunak, S., Gilbert, M. T. P., and Sicheritz-Pontén, T. (2018). Protective role of the vulture facial skin and gut microbiomes aid adaptation to scavenging. *Acta Veterinaria Scandinavica*, *60*(1), 61. https://doi.org/10.1186/s13028-018-0415-3

Zhang, F., and Lupski, J. R. (2015). Non-coding genetic variants in human disease. *Human Molecular Genetics*, *24*(R1), R102–R110. https://doi.org/10.1093/hmg/ddv259

Zhang, G. (2015). Bird sequencing project takes off. *Nature*, *522*(7554), Article 7554. https://doi.org/10.1038/522034d

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., … Wang, J. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science (New York, N.Y.)*, *346*(6215), 1311–1320. https://doi.org/10.1126/science.1251385

Zhang, P., Wu, W., Chen, Q., and Chen, M. (2019). Non-Coding RNAs and their Integrated Networks. *Journal of Integrative Bioinformatics*, *16*(3), 20190027. https://doi.org/10.1515/jib-2019-0027

Zhou, Q., Zhang, J., Bachtrog, D., An, N., Huang, Q., Jarvis, E. D., Gilbert, M. T. P., and Zhang, G. (2014). Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science*, *346*(6215), 1246338. https://doi.org/10.1126/science.1246338

Zhou, X., Lin, Q., Fang, W., and Chen, X. (2014). The complete mitochondrial genomes of sixteen ardeid birds revealing the evolutionary process of the gene rearrangements. *BMC Genomics*, *15*(1), 573. https://doi.org/10.1186/1471-2164-15-573

Zhou, Z. (2014). The Jehol Biota, an Early Cretaceous terrestrial Lagerstätte: New discoveries and implications. *National Science Review*, *1*(4), 543–559. https://doi.org/10.1093/nsr/nwu055

Zhou, Z., and Zhang, F. (2007). Mesozoic birds of China—A synoptic review. *Frontiers of Biology in China*, *2*(1), 1–14. https://doi.org/10.1007/s11515-007-0001-y

Zou, D., Tian, S., Zhang, T., Zhuoma, N., Wu, G., Wang, M., Dong, L., Rossiter, S. J., and Zhao, H. (2021). Vulture Genomes Reveal Molecular Adaptations Underlying Obligate Scavenging and Low Levels of Genetic Diversity. *Molecular Biology and Evolution*, *38*(9), 3649–3663. https://doi.org/10.1093/molbev/msab130

Chapter two – Whole genome sequencing and assembly of the Southern Ground Hornbill (*Bucorvus leadbeateri*)

2.1 Abstract

Genomics, a field experiencing rapid evolution, has been transformed by advancements in sequencing technologies. These advancements have enabled the resolution of genomes of numerous avian species. However, there is a notable lack of genomic information available for the Southern Ground Hornbill (SGH – *Bucorvus leadbeateri*), a flagship species of the African savannah biome. Here, we undertook the sequencing, assembly and annotation of the complete SGH genome. The genome was sequenced using both long-read (Pacific Bioscience - PacBio) and short-read (Illumina NovaSeq 6000) technologies. We then evaluated the performance of stand-alone long-read assemblers Canu and wtdbg2, short-read assembler ABySS and a hybrid *de novo* assembler DBG2OLC to obtain the optimal genome assembly. The hybrid approach produced the most contiguous assembly, composed of 10,518 contigs, an N50 value of 0.26 Mb and a genome size of 1.12 Gb. This assembly was further refined using contig extension, polishing, scaffolding and gap closing. The final 1.16 Gb assembly comprised of 1,674 contigs with an N50 value of 40.45 Mb. The SGH genome provides a pivotal foundation for resolving the genetic blueprint of this species, enabling the identification of genetic pathways and mechanisms utilised in their survival. This will aid in conservation efforts and ultimately help preserve the history of the SGH in the possible event of extinction.

2.2 Introduction

The field of avian genomics has witnessed a veritable explosion in the last decade, with the genomes of 3,811 avian species sequenced and assembled by the Bird 10,000 Genomes Project (B10K, 2024) and 1,601 avian genomes (www.ncbi.nlm.nih.gov) being available to date. This progress is largely attributed to the advent and evolution of next-generation sequencing technologies (Delmore and Liedvogel, 2020). Currently, the two most common modes of sequencing whole genomes include second-generation (short-read) and third-generation (long-read) sequencing (Amarasinghe *et al*., 2020; Gupta and Verma, 2019).

Second-generation sequencing technologies encompass several sequencing platforms, including Roche/454 Life Sciences, Illumina/Solexa, ABI/SOLiD, and Ion Torrent (Heather and Chain, 2016; Kchouk *et al*., 2017; Verma *et al*., 2017). The platforms allow for the mass parallelization of sequencing reactions through the use of polymerase chain reaction (PCR). This technique significantly increases the amount of DNA that can be sequenced in a single run, thereby accelerating the pace of genomic research (Heather and Chain, 2016). However,

these second-generation sequencing technologies have major drawbacks including the production of short-read lengths, which can pose challenges in resolving highly repetitive regions of complex genomes (Heather and Chain, 2016; Schatz *et al*., 2010). Additionally, the use of PCR, while beneficial in terms of increasing throughput, is both time-consuming and costly and has the potential to introduce sequencing errors and biases (Pareek *et al*., 2011).

The third-generation sequencing technologies are represented by Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). In contrast to second-generation sequencing technologies, these platforms operate on the principle of single-molecule real-time sequencing (Athanasopoulou *et al*., 2022). This results in a substantial reduction in sequencing costs, simplified sample preparation, and improved execution time. Furthermore, they have enabled the production of long reads, a notable improvement over their predecessors (Pareek *et al*., 2011; Weirather *et al*., 2017). However, these improvements come with their own set of challenges. The major drawbacks of these two technologies are the high error rates they produce, and the lower throughput achieved compared to most second-generation sequencing technologies (Lin *et al*., 2021; Rhoads and Au, 2015; Weirather *et al*., 2017). Despite these challenges, the benefits of third-generation sequencing technologies in providing a more comprehensive view of the genome structure cannot be overstated. Furthermore, with the introduction of techniques such as PacBio High Fidelity (HiFi) sequencing which combines the benefits of both short and long reads, highly accurate long reads can now be achieved (Wenger *et al*., 2019). More recently, there has been a shift towards the integration of both short- and long-read sequencing data, combining the strength of both techniques, to achieve a more effective overview of the genome structure of avian species (Rhie *et al*., 2021).

In this study, we undertook the genome sequencing of the Southern Ground Hornbill (SGH – *Bucorvus leadbeateri*) using both long- and short-read technologies to gain insight into the unique biological, behavioural, and lifestyle adaptations of this bird. We benchmarked the performance of several assembly algorithms and tools (short- and long-read-only and hybrid *de novo* assemblers) in producing the most contiguous, complete, and accurate genome assembly of the SGH. The high-quality draft genome was subsequently structurally and functionally annotated. The availability of this genome will aid in improving our current understanding of this bird at a genomic level and provide a pivotal foundation for comparative genomic analysis with other hornbill species.

2.3 Methods and materials

2.3.1 Sample collection and DNA extraction

A blood sample (~0.5 ml) was collected from a captive female SGH at the Montecasino Bird Gardens (Johannesburg, South Africa) in collaboration with the Mabula Ground Hornbill Project. A female SGH was chosen to access both the Z and W sex chromosomes during genome assembly. As the SGH is classified as threatened or protected in terms of Section 56 of the National Environmental Management: Biodiversity Act (Act no. 10 of 2004) the following clearance certificates and permits were obtained prior to sample collection: (1) Ethics clearance from the University of the Witwatersrand Animal Ethics Screening Committee (Clearance number: 2020/09/03B), (2) Clearance from the Department of Agriculture, Land Reform and Rural Development (DALRRD), Section 20 of the Animal Disease Act, 1984 (Act no. 35 of 1984) to work on biological material (Reference number: 12/11/1/1/20 (1775JD)) and (3) Clearance from the Department of Environmental Affairs Threatened Or Protected Species (TOPS) permit (Certificate number: 29480). The latter was renewed via the School of Animal, Plant and Environmental Sciences at the University of the Witwatersrand.

Total genomic DNA was extracted using the phenol-chloroform-isoamyl alcohol approach. Pelleted cells from the blood sample were suspended in a lysis buffer composed of STE buffer (0.1 M Sodium Chloride, 0.05 M Tris-HCl and 0.1 M EDTA), 20 mg/ml Proteinase K, 10 mg/ml RNase and 10% Sodium dodecyl sulphate (SDS) and incubated at 50 °C overnight. Organic extraction of the nucleic acids was performed using phenol-chloroform-isoamyl alcohol (25:24:1), which promoted the partitioning of cellular debris and lipids in the organic phase and the DNA in the aqueous phase. The two phases were separated via centrifugation and the aqueous phase was retained. The DNA was precipitated with 95% ethanol and 3 M sodium acetate and incubated at -20 °C overnight. The precipitated DNA was pelleted via centrifugation and excess salts present in the sample were washed with 70% ice-cold ethanol. Residual ethanol was evaporated from the sample and the DNA was eluted in TE buffer (0.01M Tris-HCl and 0.001M EDTA). DNA concentration and quality were evaluated using a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA) and gel electrophoresis using a 1% agarose gel.

The DNA was sent to Molecular Research LP (www.mrdnalab.com, Shallowater, Texas, USA) for sequencing, where DNA concentration and quality were re-evaluated using Qubit ® dsDNA

HS Assay kit (Life Technologies, California, USA) and a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA). The sample quality was improved using the DNeasy PowerClean Pro cleanup kit (Qiagen, Maryland, USA). The size of the genomic DNA for long-read sequencing was determined using electrophoresis with the E-Gel SizeSelect 2% Agarose gel (Invitrogen, California, USA).

2.3.2 Library preparation and sequencing

2.3.2.1 PacBio Sequel library preparation and sequencing

The gDNA was sheared using a Covaris G-tube (Covaris Inc., Massachusetts, USA) and the average size of the sample was evaluated using the Agilent 2100 Bioanalyzer (Agilent Technologies, California USA). Approximately 1 µg of genomic DNA was used for library preparation using the SMRTbell Express Template Prep kit 2.0 (Pacific Bioscience, California, USA). During library preparation, the DNA was fragmented (sheared), followed by the removal of single-strand overhangs and the repair of any DNA damage present. Lastly, barcode adapters were ligated. Following library preparation, DNA fragments that ranged between 6 to 10 kilobases (kb) were selected using the Blue Pippin automated size-selection instrument (Sage Science, Massachusetts, USA). The final concentration of the library was evaluated using the Qubit ® dsDNA HS Assay kit (Life Technologies, California, USA) and the average size of the library was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, California USA). The library was then sequenced using the PacBio Sequel (Pacific Bioscience, California, USA), with the paired-end 10-hour sequencing protocol. This approach produced both Continuous Long Reads (CLR) and HiFi reads. CLR reads are produced by sequencing a SMRTbell template with a large insert size. The large insert size limits the polymerase to a single pass of the template molecule, which leads to a high error rate (Logsdon *et al*., 2020). In contrast, HiFi reads are generated through circular consensus sequencing. This method also uses the SMRTbell template but with a smaller insert size (Wenger *et al*., 2019). The advantage of the smaller insert size permits the polymerase to make multiple passes over the same template strand. This repetition significantly improves the accuracy of the sequenced reads (Wenger *et al*., 2019). In this study, the HiFi read dataset produced a low coverage, thus, an amalgamated dataset comprised of 5 copies of HiFi reads (5X HiFi) and one copy of the CLR (1X CLR) reads was generated and subsequently used for genome assembly.

2.3.2.2 Illumina library preparation and sequencing

Approximately 50 ng of genomic DNA was used for library preparation using the Illumina DNA Prep (M) Tagmentation Library Preparation kit (Illumina, California, USA). During library preparation, the genomic DNA sample was diluted, fragmented and tagged with adapters. These adaptors added unique barcodes to the fragmented DNA during a limited cycle PCR, which was conducted to avoid over-amplification and biases in the reaction library (Aird *et al*., 2011; Head *et al*., 2014). Following library preparation, the final concentration of the library was evaluated using the Qubit ® dsDNA HS Assay kit (Life Technologies, California, USA) and the average size of the library was determined using the Agilent 2100 Bioanalyzer (Agilent Technologies, California USA). The library was then sequenced using the Illumina NovaSeq 6000 (Illumina, California, USA) following a paired-end (PE) 300-cycle protocol.

2.3.3 Raw data quality control and processing

The quality of the raw PacBio and Illumina reads was evaluated using FastQC v 0.11.9 (Andrew, 2015). Adaptors and low-quality bases (Phred score (Q) < 25) in the PE Illumina reads were trimmed using the script fastq_quality_trimmer from the FastX_toolkit v 0.0.14 (Hannon Lab, n.d.). The trimmed reads were then re-evaluated using FastQC. Assessment of the quality of raw sequenced data is crucial prior to assembly as the presence of poor-quality reads, ambiguous base calling, adapter contamination and biases in data may affect downstream analysis resulting in misinterpretations of underlying genetic information (Angel *et al*., 2018).

2.3.4 *De novo* genome assembly

In order to achieve the best possible genome assembly, the performance of four assembly pipelines with different algorithms was evaluated: 1) Assembly By Short Sequences (ABySS) (Jackman *et al*., 2017) for the Illumina short reads, 2) Canu (Koren *et al*., 2017) and 3) wtdbg2 (Ruan and Li, 2020) for the PacBio long reads and 4) a hybrid genome assembler DBG2OLC (Ye *et al*., 2016) that used a combination of the Illumina and PacBio reads (Figure 2.1). Quast-LG v 5.2.0 (Gurevich *et al*., 2013) was systematically used at each step of the assembly process to evaluate the performance of the various assembly tools. Measures of performance included the number of contigs generated, the length of the largest contig achieved in the assembly and the N50 value. The N50 measure is the length of the shortest contig for which equal length or larger contigs cover 50% of the assembly (Mikheenko *et al*., 2018). Furthermore, the genome fraction (percentage of bases in the draft genome that aligns with a reference genome) was

calculated using the Abyssinian Ground Hornbill (AGH – *Bucorvus abyssinicus*; GCA_009769605.1; Rhie *et al*., 2021) as a reference. Lastly, the average number of gaps/uncalled bases (N's) were taken into consideration to evaluate the performance of the gap filling tools (Mikheenko *et al*., 2018).

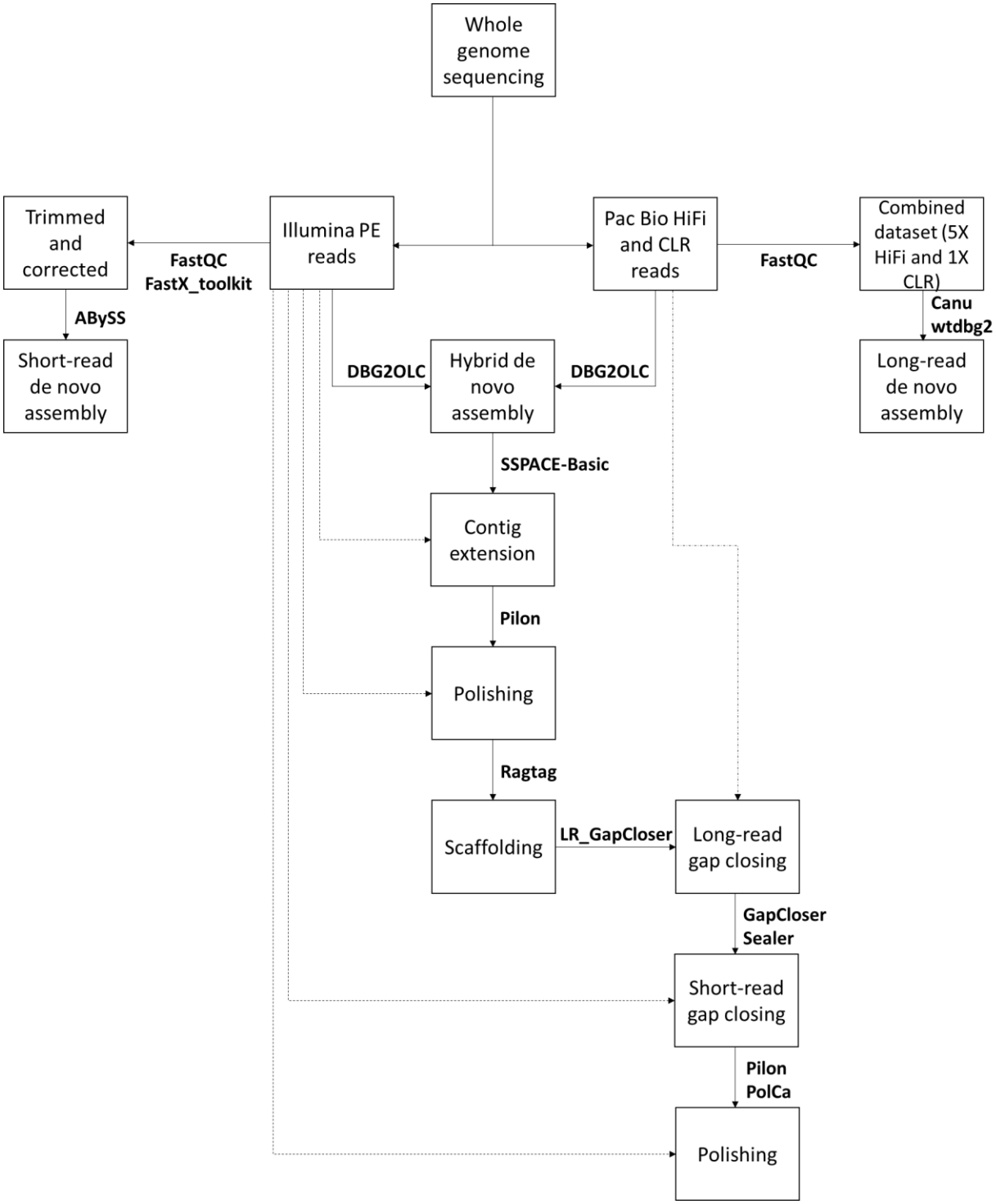


Figure 2.1: Workflow of the SGH *de novo* genome assembly approach.

2.3.4.1 Long-read de novo genome assembly

The amalgamated PacBio dataset (5X HiFi and 1X CLR reads) was assembled using Canu v 2.3 (Koren *et al*., 2017) with default parameters and an estimated genome size of 1.2 Gb. In brief, overlaps between the reads were detected by the MinHash alignment process (MHAP) (Berlin *et al*., 2015) to identify discrepancies and correct them using a consensus sequence. The reads were then trimmed to remove low-quality sequences and remaining adapters and then assembled into contigs using the best overlap-layout-consensus (OLC) assembly graph (Koren *et al*., 2017). In the OLC assembly approach, overlaps between all the reads are first identified and then a layout of the assembly graph is created using all the reads and the overlaps information. Lastly, a consensus sequence is inferred from the assembly graph (Li *et al*., 2012). Similarly, the amalgamated PacBio dataset was assembled using wtdbg2 (Ruan and Li, 2020), with default settings. In contrast to Canu, wtdbg2 assembled the reads without prior error correction and used a fuzzy De Bruijn assembly graph to generate contigs (Ruan and Li, 2020). The De Bruijn assembly approach segments the reads into shorter k-mers, which are substrings of fixed length. The k-mers are then used to create an assembly graph from which the genome sequence is inferred (Li *et al*., 2012). Quast-LG was used to determine the optimal long-read assembly produced from the two assemblers.

2.3.4.2 Short-read de novo genome assembly

The Illumina short reads were assembled using ABySS v 2.0 (Jackman *et al*., 2017), with default settings and varying k-mer lengths. K-mers are crucial for the construction of De Bruijn assembly graphs which is the strategy utilised by ABySS for contig generation (Jackman *et al*., 2017; Simpson *et al*., 2009). KmerGenie v 1.7051 (Chikhi and Medvedev, 2014) was employed to estimate the best k-mer length for the quality filtered and trimmed Illumina reads. An optimal k-mer length of 31 was determined. To optimize the short-read assembly, ABySS was evaluated using the optimal and upper and lower bound k-mer lengths (21, 30-32 and 41).

2.3.4.3 Hybrid de novo genome assembly

The hybrid *de novo* assembler DBG2OLC was used to produce an assembly using a combination of the raw PacBio long reads and Illumina short reads. DBG2OLC functions in an iterative process comprised of three steps: a De Bruijn graph assembly, construction of the best overlap-layout assembly graph and consensus calling (Ye *et al*., 2016). In brief, a De Bruijn graph was first constructed with the Illumina reads using Sparse Assembler (Ye *et al*., 2016,

2012). The parameters k-mer length [15 – 127], gap size (g - number of intermediate k-mers skipped) [1 – 25], Node Coverage Threshold (NodeCovTh – the k-mer coverage threshold used to filter out low coverage or spurious k-mers representative of sequencing errors or artifacts introduced by non-biological processes such as chemical reactions during sequencing) [0 – 16] and the Edge Coverage Threshold (EdgeCovTh – coverage threshold used to filter out low coverage and spurious links which are edges in the assembly graph representative of sequencing errors or artifacts) [0 – 16] were optimised within the recommended range (Ye *et al*., 2012). Quast-LG was used to identify the optimal parameter values for the Sparse Assembler step. The final Sparse Assembler step was run using the following parameters: k-mer size = 31, gap size = 13, NodeCovTh = 1, EdgeCovTh = 0 and a default LD (loading the saved compressed or anchored reads) = 1.

The contigs generated by Sparse Assembler were mapped against the PacBio long-read dataset and used to correct errors and trim reads with structural errors (Ye *et al*., 2016). During this stage, the long reads were compressed into contig identifiers to reduce the computational costs required to process long reads (Ye *et al*., 2016). The error-corrected reads were then used to construct the best overlap assembly graph to determine the correct order and orientation of the reads in the final assembly (Ye *et al*., 2016). During this overlap and layout step, the parameters Adaptive Threshold (AdaptiveTh – the threshold for matching the adaptive k-mers) [0.001 – 0.02], Kmer Coverage (KmerCov – the threshold for matching fixed k-mers) [2 – 10] and Minimum Overlap (MinOverlap – minimum overlap score between long-read pairs) [10 – 150] were optimised within the recommended range for PacBio data between a coverage of 10X to 100X. Quast-LG (Mikheenko *et al*., 2018) was used to identify the optimal parameter values for the overlap and layout step. The final overlap and layout step was run using the following parameters: AdaptiveTh = 0.003, KmerCovTh = 2, MinOverlap = 21, a default k-mer = 17 and RemoveChimera (remove structural errors) = 1.

Lastly, consensus calling using BLASR v 5.3.5 (basic local alignment with successive refinement) (Chaisson and Tesler, 2012) and Sparc (Ye and Ma, 2016) was conducted to polish the final assembly using default settings. In brief, the long reads were first uncompressed and chained together based on the best overlaps detected and the contigs generated using the Illumina reads were used to fill any gaps present in the long reads. BLASR and Sparc were then used to align the reads and calculate the polished assembly using an efficient sparse k-mer graph (Ye *et al*., 2016).

2.3.5 Refining the draft SGH genome assembly

The best performing draft SGH genome assembly (produced using DBG2OLC) was improved in an iterative process using multiple tools (Figure 2.1). Quast-LG was used to determine the assembly statistics and identify the optimal assembly improvement tool suited for the data. Initial elongation of the pre-assembled contigs was performed to improve sequence contiguity and reduce the number of contigs generated through the process of contig extension (Boetzer *et al*., 2011; Deng and Delwart, 2021). Contig extension was performed by evaluating the performance of both long- and short-read tools. SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension Long Read (SSPACE-LR) (Boetzer and Pirovano, 2014) was run with PacBio HiFi and CLR reads with default parameters. SSPACE-Basic was run with the Illumina reads with an average insert size of 425 base pairs and a length variation of 0.5. In brief, raw Illumina and PacBio reads were mapped against the pre-assembled genome using Bowtie (Langmead *et al*., 2009) and BLASR, respectively. Overlaps between the ends of the contigs were identified, from which the best overlaps were used to link contigs and extend them (Boetzer *et al*., 2011; Boetzer and Pirovano, 2014).

Reads generated from long-read sequencing technologies tend to have a higher error rate compared to reads generated using short-read sequencing technologies (Jung *et al*., 2020). This increases the probability of base calling errors, small insertions/deletions (INDELs), gaps, alignment discontinuity and ambiguities as a result of repetitive regions (Jung *et al*., 2020; Zimin and Salzberg, 2020). Thus, to improve the accuracy of the assembled genome, polishing of the contig extended assembly was conducted using Racon for long erroneous reads (Vaser *et al*., 2017) and Pilon for short-read sequenced data (Walker *et al*., 2014). In brief, the PacBio HiFi reads were aligned against the assembled contigs using Minimap (Li, 2016) for Racon, while Illumina reads were aligned using BWA (Li, 2014) for Pilon. The aligned reads were used to identify errors and discrepancies between the reads which were subsequently corrected. The error-corrected reads were then used to generate updated consensus sequences (Vaser *et al*., 2017; Walker *et al*., 2014). Polishing using Pilon was conducted iteratively (ten rounds) to progressively improve the accuracy of the assembled genome.

In order to orientate and order the pre-assembled contigs of the polished SGH draft genome assembly and further improve sequence contiguity, scaffolding (generating longer fragments of sequences with stretches of N nucleotides representing gaps where the nucleotide sequences were unresolved) was conducted using the AGH genome assembly (GCA_009769605.1) as

reference (Luo *et al*., 2021; Shieh *et al*., 2020). Additional data regarding the chromosomal structure of the SGH, such as optical and linkage maps are currently not available. Therefore, a reference-guided scaffolding approach was conducted in comparison to reference-free scaffolding (Alonge *et al*., 2019). This was done to enhance the accuracy and completeness of the SGH draft genome assembly (Alonge *et al*., 2019). Scaffolding of the polished reads was conducted using RagTag (Alonge *et al*., 2022) and Chromosome_Scaffolder (Zimin *et al*., 2013) independently against the AGH genome. In brief, the pre-assembled contigs were mapped against the AGH genome using Minimap2 (Li, 2018) and Nucmer (Marçais *et al*., 2018), respectively. The purpose of this mapping was to identify the chromosome in the AGH genome that had the most alignments with the contigs. Once these regions of alignment on the AGH chromosomes were identified, they were used to infer the order and orientation of the contigs. These inferences were then applied to arrange the contigs within the SGH draft genome assembly (Luo *et al*., 2012).

Scaffolding of a genome has the drawback of leaving variable amounts of unknown sequence (gaps) between ordered contigs within a scaffold where no overlaps were found (Kammonen *et al*., 2019). Gaps can be a direct result of repetitive sequences, complex genomic architecture and the performance of sequencing platforms and assembly tools (Xu *et al*., 2020). Thus, gap closing was used to resolve the unknown sequences with the use of the raw sequence data. For the SGH, both long (PacBio) and short (Illumina) read gap closing was performed. Gap closing using the independent PacBio CLR and HiFi reads was performed using LR_Gapcloser and TGS_GapCloser (Xu *et al*., 2019, 2020). LR_Gapcloser was also run with a combination of both the CLR and HiFi reads. Gap closing with a combination of the PacBio CLR and HiFi dataset was further evaluated using Samba from MaSurCa (Zimin and Salzberg, 2022). In brief, LR_Gapcloser functions on the basis of a tiling path created by aligning fragmented raw PacBio reads against the assembled scaffolds using BWA-MEM (Li, 2014). TGS_GapCloser and Samba in contrast aligned the raw unfragmented PacBio reads against the scaffolds using Minimap 2 (Li, 2016). Based on the alignment results, gaps with sequence information available from the long-read data were extended and filled to create a refined consensus sequence.

Gap closing for the Illumina PE reads was performed with Gap2Seq (Salmela and Tomescu, 2019), GapCloser (Luo *et al*., 2012) and Sealer (Paulino *et al*., 2015). In contrast to long-read gap closing, these short-read gap closing tools use the graph-based (usually a de Bruijn graph)

method to resolve gaps (Walve *et al*., 2017). The initial short-read assembly graph is used in combination with the short-read sequenced data to identify regions of no overlaps indicating gaps. Paths within the assembly graph are then explored with a range of k-mer lengths to identify sequences that can successfully extend and fill the gaps and a consensus sequence is generated (Paulino *et al*., 2015; Walve *et al*., 2017).

This was followed by a final round of polishing using five iterations of Pilon and then five iterations of PolCa (Zimin and Salzberg, 2020) to identify and correct any errors resulting from gap filling using the raw PacBio and Illumina reads. Alignment of the assembled reads to the Illumina reads was conducted using BWA for Pilon and BWA-MEM for PolCa (Li, 2014; Li and Durbin, 2009; Walker *et al*., 2014; Zimin and Salzberg, 2020), respectively.

2.3.6 Genome annotation

Genome annotation was performed using the annotation tool Modular Open-Source Genome Annotator (MOSGA) via an iterative process. First, repetitive elements such as low-complexity DNA sequences and interspersed repeats were identified and classified using RepeatMasker v 4.1.1 (Smit *et al*., 1996) by searching against the Dfam (Wheeler and Eddy, 2013) database using nhmmer v 3.3.2 (Wheeler and Eddy, 2013) from HMMER v 3.1 (Eddy, 2011). Instead of removing these identified repetitive elements from the genome using hard-masking (replacing with Ns), the nucleotide content of the repetitive elements was conserved using soft-masking (replacing with lowercase characters) with the *B. abyssinicus* model as a reference (Smit *et al*., 1996; Tarailo-Graovac and Chen, 2009).

Ab initio gene prediction, gene prediction based on statistical models trained using known proteins and gene models rather than experimental evidence, was performed on the soft-masked SGH genome using Augustus (Scalzitti *et al*., 2020) trained with the chicken (*Gallus gallus* – GalGal4) gene model. The chicken gene model was used to train Augustus as it shares common ancestry with other avian species, resulting in the conservation of many genomic features such as regulatory regions, coding and non-coding elements across species (Burt, 2005). Furthermore, it is one of the most well-resolved avian gene models (Cheng and Burt, 2018; Zhang *et al*., 2014). MOSGA also takes the results generated by Benchmarking Universal Single-Copy Orthologs (BUSCO), which assesses the quality of gene models in a genome using the Aves lineage dataset aves_odb10 to identify genes common to all organisms (Simão *et al*., 2015; Waterhouse *et al*., 2018). It then uses these results to retrain Augustus and adjust

its parameters. This improves the accuracy of Augustus in predicting the locations and structures of genes and proteins in the genome (Martin *et al*., 2021). Furthermore, this process enables Augustus to identify statistical patterns linked to the gene coding regions in other bird species, such as the SGH, thereby improving the accuracy of gene and protein predictions across multiple species (Stanke *et al*., 2006).

Transfer RNAs (tRNAs), which are critical in protein synthesis, were predicted using tRNAscan-SE v 2.0 (Chan *et al*., 2021) with default parameters. This program functions using covariance models capturing both the primary sequence and characteristic cloverleaf secondary structure of tRNAs to scan the input sequence for potential tRNA genes (Chan and Lowe, 2019). As tRNAs are conserved across evolutionary lineages, tRNAscan-SE v 2.0 was trained using the Eukaryotic domain dataset, which contains a gene set model comprised of 110 genera and 155 genomes (Chan *et al*., 2021).

Ribosomal RNA (rRNA) genes, which are crucial components of the ribosome responsible for protein synthesis, were predicted using Basic Rapid Ribosomal RNA Predictor (Barrnap) v 0.9 (Seemann, 2013) set to identify conserved eukaryotic 5S, 5.8S, 28S and 18S rRNA with a minimum E-value of 10e-6. In brief, Barrnap used nhmmer (Wheeler and Eddy, 2013) from HMMER v 3.1 (Eddy, 2011) to search the input sequence for Hidden Markov Model (HMM) profiles, built specifically for the eukaryotic rRNA families, to identify regions that matched the conserved features of the rRNA genes (Seemann, 2013).

The completeness of the final draft genome and the quality of the predicted gene models were also assessed using BUSCO with default parameters, using the Aves lineage dataset aves_odb10. This dataset is composed of 8,338 genes that are conserved across the avian lineage and that only occur as single-copy orthologues (Simão *et al*., 2015; Waterhouse *et al*., 2018). The completeness of the final SGH draft assembly and annotations was evaluated based on the presence of the database genes on the SGH genome.

Predicted proteins in the SGH were functionally annotated using evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG) mapper v. 2 (Cantalapiedra *et al*., 2021) against the eggNOG 5.0 database (Huerta-Cepas *et al*., 2019), with default parameters. eggNOG functionally annotates the predicted genes using Conserved Orthologous Groups (COGs) categories (Tatusov *et al*., 2000), Gene ontology (GO) terms (Ashburner *et al*., 2000; Gene Ontology Consortium *et al*., 2023), Kyoto Encyclopedia of Genes and Genomes (KEGG)

pathways (Kanehisa and Goto, 2000), Simple Modular Architecture Research Tool (SMART)/PFAM domains (Schultz *et al*., 1998; Sonnhammer *et al*., 1997), Carbohydrate-active enzymes (CAZy) (Cantarel *et al*., 2009) and KEGG modules (Kanehisa *et al*., 2008).

2.4 Results

2.4.1 Raw sequence data assessment

The SGH genome was initially sequenced using the PacBio Sequel platform which resulted in the generation of a total of 5,403,320 reads with an average sequence length of 7,847 bp (181 – 18,052 bp) for the HiFi reads and 6,949 bp (51 – 100,084 bp) for the CLR reads (Table 2.1). A genome coverage (calculated on the basis of final assembly size) of 30.51x was achieved for the CLR reads, while a lower-than-expected genome coverage (2.22x) was obtained using the HiFi sequencing configuration. Given the higher accuracy of HiFi reads, in order to compensate for its lower genome coverage, an amalgamated dataset containing five copies of the HiFi (5X HiFi) dataset with one copy of the CLR reads was generated, resulting in a genome coverage of 41.6 x for the PacBio sequencing.

To further improve the genome coverage as well as base calling accuracy, the SGH genome was also sequenced using the Illumina NovaSeq 6000 platform using the paired-end (2x 150 bp) configuration. This resulted in the generation of a total of 45,146,317 PE reads with an average sequence length of 290 bp (2x 145 bp), with individual reads ranging between 35 – 151 bp (Table 2.1). The total length of the short reads was 13.07 Gb, which provided a final genome coverage of 11.27x. Quality analysis with FastQC indicated the presence of low-quality base sequences ($Q < 25$) between the read positions of 150 – 151 bp. The Illumina reads were thus trimmed to remove low-quality base sequences and adapters which resulted in the removal of 969 read pairs and ~ 12.5 Mb of poor-quality sequence data.

Table 2.1: Genome sequencing results of the PacBio and Illumina reads

Assembly	Illumina (PE)	CLR reads	HiFi reads
Number of reads	45,146,317	5,088,541	327,323
Average read length (bp)	290 (2x 145)	6,949	7,847
Total length of reads (bp)	13,071,275,633	35,359,661,685	2,568,378,704
GC (%)	44.50	43.72	43.31
N50 (bp)	302	7,859	7,856
N90 (bp)	296	5,975	6,812
Number of N's per 100 kb	5.88	0.00	0.00
Genome coverage	11.27	30.51	2.22

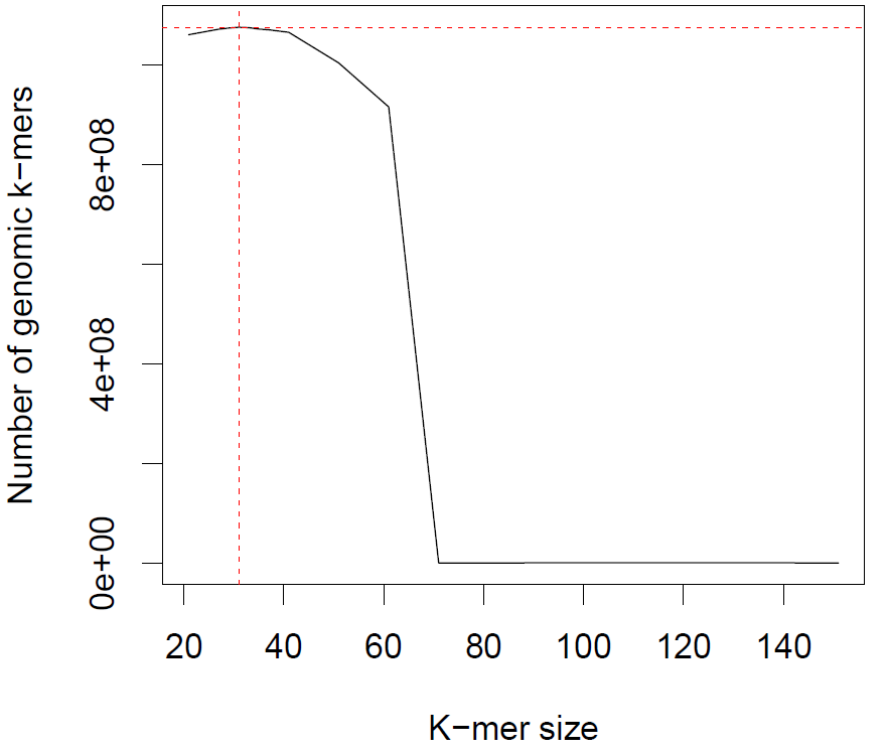
2.4.2 *De novo* genome assemblies

To achieve an accurate, contiguous and complete draft genome of the SGH, the Illumina PE reads and PacBio reads (1x CLR + 5x HiFi data combination) were individually assembled using a stand-alone short-read (ABySS) and long-read (Canu and wtdbg2), respectively.

2.4.2.1 Short-read de novo assembly

KmerGenie (Chikhi and Medvedev, 2014) was first used to determine the optimal k-mer length for the Illumina PE reads at intervals of ten. This indicated that the abundance of k-mers, which are sequences of length “k”, was highest between a k-mer size of 20 to 40 (Figure 2.2a). Specifically, there were over nine hundred million k-mers within this range. However, as the k-mer size increased beyond this range, the abundance began to decrease. To pinpoint the exact k-mer length that yielded the highest abundance within the 20 to 40 range, the interval was narrowed down to increments of five. This revealed that the peak abundance of k-mers, exceeding 1,075,000,000, occurred at a k-mer size of 31 (Figure 2.2b). This suggests that a k-mer size of 31 was optimal for this dataset.

a)



b)

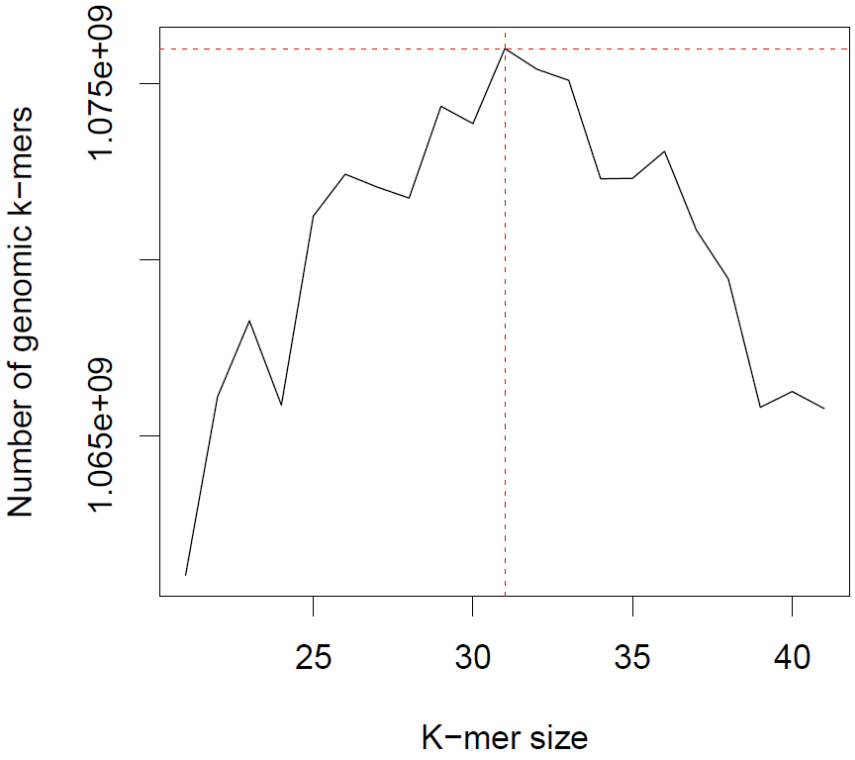


Figure 2.2: Identification of optimal k-mer length for the short-read Illumina dataset. The abundance of k-mers across the genome was identified at a k-mer length interval of a) ten between 20 to 140 and then b) five between 25 to 40.

The ABySS assembler was assessed using a selection of k-mer lengths (21, 30 – 32 and 41) centred around the optimal k-mer length identified. Comparisons of the assembly statistics of the ABySS assemblies indicated that a k-mer value of 21 resulted in the least contiguous assembly, with the highest number of contigs, the lowest N50 value and the lowest genome fraction when mapped against the AGH genome (Table 2.2). Similar decreases in the contiguity of the genome assemblies and the N50 values were observed with a k-mer length of 41. The AbySS assemblies with k-mer lengths of 30 to 32 achieved genome fractions > 89%, indicating that a greater proportion of bases were covered by these assemblies when mapped against the AGH genome. Amongst these, the assembly with the k-mer length 32 achieved the highest genome fraction when mapped against the AGH genome, however, on average 848,772 more contigs were generated and the N50 value was on average 606 bp smaller than the assemblies achieved with a k-mer length of 30 and 31. A comparison of the assemblies generated with k-mer lengths 30 and 31 indicated that a k-mer length of 30 produced an assembly with a lower genome fraction (0.11% less) than a k-mer length of 31 when mapped against the AGH genome, but the former assembly was more contiguous (78,295 fewer contigs) and had a higher N50 value (38 bp greater). Thus, the *de novo* short-read assembly with a k-mer length of 30 was the optimal short-read assembly achieved.

Table 2.2: ABySS assembly statistics across a range of k-mer lengths evaluated

Assembly	21	30	31	32	41
Number of contigs	10,564,836	3,178,635	3,256,930	4,066,554	3,357,900
Largest contig (bp)	2,238	24,267	25,583	19,367	24,185
Total length of genome (bp)	1,236,497,480	1,173,056,729	1,180,238,567	1,252,622,050	1,188,146,842
Reference genome length (bp)	1,132,614,622	1,132,614,622	1,132,614,622	1,132,614,622	1,132,614,622
GC (%)	42.62	43.32	43.36	43.66	43.40
Reference GC (%)	43.30	43.30	43.30	43.30	43.30
N50 (bp)	166	1,790	1,752	1,165	1,703
N90 (bp)	49	115	106	81	98
L50 (bp)	2,263,205	178,856	183,258	276,437	189,032
L90 (bp)	7,525,772	1,056,744	1,108,539	1,846,000	1,172,619
Number of misassemblies	32	111	106	71	98
Number of misassembled contigs	32	111	106	71	98
AGH genome fraction (%)	41.71	89.08	89.17	89.28	89.25
Number of N's per 100 kb	0.01	10.51	10.25	5.68	9.65

2.4.2.2 Long-read de novo assembly

Comparison of the long-read assemblies generated by Canu and wtdbg2 (Koren *et al*., 2017; Ruan and Li, 2020) with the amalgamated PacBio dataset showed that the assembly generated by Canu had a genome fraction that was ~ 5% greater than the wtdbg2 assembly when mapped against the AGH genome (Table 2.3). The assembly generated with Canu also comprised 28,857 fewer contigs, with the largest contig being 0.22 Mb larger than that of the wtdbg2 assembly. The wtdbg2 assembly further generated a genome that was 0.16 Gb smaller compared to that produced using Canu. The substantial difference in the performance of the long-read assemblers could be attributed to their underlying objectives. Wtdbg2 aims to produce the fastest possible assembly by minimizing computational costs via a fuzzy de Bruijn graph approach (Ruan and Li, 2020). In contrast, Canu aims to produce the most accurate genome assembly possible. To achieve this, Canu incorporates multiple rounds of error correction and uses the OLC approach in a greedy process, making optimal choices at each

step of the overlap assembly (Koren *et al*., 2017). These differences could have enabled Canu to produce the optimal long-read assembly. The optimal long- and short-read assemblies by Canu and ABySS with a k-mer length of 30 both achieved a genome fraction over 89% when mapped against the AGH genome (Table 2.3). However, the long-read assembly by Canu was more contiguous as it generated 3,148,618 fewer contigs and had an N50 value that was 49.99 kb greater in size than the optimal ABySS assembly.

Table 2.3: Canu and wtdbg2 long-read assembly statistics compared to optimal short-read assembly

Assembly	**Canu**	**wtdbg2**	**Optimal ABySS assembly k-mer length 30**
Number of contigs	30,017	58,874	3,178,635
Largest contig (bp)	400,499	178,876	24,267
Total length of genome (bp)	1,039,480,876	973,295,628	1,173,056,729
Reference genome length	1,132,614,622	1,132,614,622	1,132,614,622
GC (%)	43.37	43.06	43.32
Reference GC (%)	43.30	43.30	43.30
N50 (bp)	51,777	20,283	1,790
N90 (bp)	15,489	8,192	115
L50 (bp)	6,065	15,140	178,856
L90 (bp)	20,276	44,975	1,056,744
Number of misassemblies	2,153	868	111
Number of misassembled contigs	1,667	776	111
AGH genome fraction (%)	89.17	84.25	89.08
Number of N's per 100 kb	0.00	0.00	10.51

2.4.2.3 Hybrid de novo assembly

The hybrid genome assembler, DBG2OLC (Ye *et al*., 2016), combines both short-read and long-read data to generate a consensus assembly. Assembly parameter optimization was performed to fine-tune the parameters of the first two steps of DBG2OLC (assembly of Illumina PE read data with SparseAssembler and the overlap and layout step) to achieve the best draft assembly. Parameters evaluated during the SparseAssembler step included k-mer length, gap size, NodeCovTh and EdgeCovTh. The latter two parameters filter out low coverage and spurious k-mers and links during assembly, respectively as they could be indicative of sequencing errors or artifacts (Ye *et al*., 2012). A total of five k-mer lengths (15,

21, 31, 51 and 71) were evaluated (Appendix Table 1). A k-mer length of 71 produced the poorest assembly with the greatest number of contigs (6,326,767), the smallest maximum contig size (15.13 kb) and the lowest N50 value (326 bp) compared to the other k-mer lengths. Furthermore, when mapped against the AGH genome, a genome fraction of only 78.68% was obtained, while > 90% was achieved for the other k-mer lengths. No difference in the key assembly metrics was observed for any of the other k-mer lengths. As such, a k-mer length of 31 was used as the k-mer length for the final run of the SparseAssembly step as it was also the optimal k-mer length determined for the Illumina reads using KmerGenie.

A total of ten distinct gap sizes (1, 5, 7, 10, 11, 12, 13, 14, 15, 25) were evaluated for their effect on SparseAssembler. The gap size of 25 produced the most contiguous assembly, with the lowest number of contigs (1,057,994) and the highest N50 value (2.91 kb) (Appendix Table 1). However, this assembly did produce the lowest genome length (1.06 Gb) and genome fraction (89.74%) compared to the lower gap sizes. In comparison gap sizes between 10 and 15 produced the highest AGH genome coverage fractions (> 90.10%) with the next highest N50 values (2.59 – 2.80 kb). As a gap size of 13 produced the highest genome fraction (90.30%) it was used for the final run of SparseAssembler.

A total of five values each (0, 1, 2, 5 and 10) were evaluated for the NodeCovTh and EdgeCovTh parameters. While a change in the EdgeCovTh had no impact on the assembly produced, a NodeCovTh value other than 1 resulted in weaker assemblies (Appendix Table 1). These assemblies were less contiguous, with the highest number of contigs produced (> 2,005,420), lower N50 values (< 1,292 bp) and lowest AGH genome coverage fraction (< 84.46%) compared to a NodeCovTh of 1 (contigs: 1,448,085; N50: 2.74 kb; genome fraction: 90.30%). As such a default EdgeCovTh of 0 and NodeCovTh of 1 were used for the SparseAssembler step.

Following the completion of the SparseAssembler step of DBG2OLC, the parameters for the overlap and layout step were also fine-tuned. First, a total of nine AdaptiveTh (0.0001, 0.001 – 0.006, 0.01 and 0.02) values were assessed. While all nine AdaptiveTh values produced an AGH genome coverage fraction of ~95%, AdaptiveTh values > 0.006 resulted in less contiguous genome assemblies (Appendix Table 1), with >11,058 contigs and the lowest N50 values (<250 kb), while AdaptiveTh values < 0.001 also resulted in N50 values below 250 kb. AdaptiveTh values between 0.002 to 0.005 performed the best with 0.003 producing 10,543

contigs, with the highest N50 value (~259 kb). As such for the final overlap and layout step, an AdaptiveTh value of 0.003 was used.

A standard k-mer length of 17 was used for the overlap and layout step as suggested by the DBG2OLC manual, however, three KmerCov values (2, 3 and 5) were tested to evaluate the impact of KmerCov on the assembly. While all three KmerCov values performed well, with an AGH genome coverage fraction of > 95.40%, the KmerCov value of 2 produced the most contiguous assembly, with the lowest number of contigs (10,543 contigs) and the highest N50 (0.26 Mb) compared to the other KmerCov values (Appendix Table 1). As such a KmerCov of 2 was used for the final overlap and layout step.

The last parameter evaluated for the overlap and layout step was the MinOverlap value for which four values (19 – 22) were evaluated. The lowest MinOverlap value (19) resulted in a poor assembly, with an AGH genome coverage fraction of only 80.72% compared to the values of 20 – 22 which attained AGH genome coverage fractions > 95.40% (Appendix Table 1). Of these MinOverlap values, 20 produced an assembly with the highest N50 value (~259 kb), however, a value of 21 produced the lowest number of contigs (10,518) and the highest genome fraction (95.47%). As such, as a MinOverlap value of 21 performed slightly better than the suggested default of 20, it was used for the overlap and layout step.

DBG2OLC hybrid assembly with optimised parameters resulted in a draft hybrid genome assembly comprised of 10,518 contigs (largest contig: 2.06 Mb), with an N50 value of 0.26 Mb (Table 2.4). The total size of the assembly was 1.12 Gb (16.53 Mb smaller than the AGH genome), with a genome coverage fraction of 95.48% when mapped against the AGH genome. The DBG2OLC assembly produced 19,499 fewer contigs compared to the optimal long-read assembly by Canu (Table 2.3). This assembly was also more contiguous with the largest contig being 1.66 Mb greater in size and the N50 value being 0.21 Mb larger than the optimal long-read assembly. Lastly, the genome fraction of the DBG2OLC assembly was 6.30% greater than the optimal long-read assembly when mapped against the AGH genome. This indicated that the DBG2OLC assembly outperformed the standalone short- and long-read assemblies in terms of contiguity and completeness (in comparison to the AGH genome) and this assembly was therefore used for subsequent analysis.

Table 2.4: Assembly statistics of the SGH draft genome assembly per improvement stage

Assembly	DBG2OLC assembly	Contig extension	Polishing	Scaffolding	Long-read gap closing	Short-read gap closing	Polishing
Number of contigs	10,518	10,513	10,513	1,674	1,674	1,674	1,674
Largest contig (bp)	2,058,996	2,058,996	2,058,332	138,051,206	138,053,805	138,051,936	137,990,494
Total length of genome (bp)	1,116,083,548	1,116,441,123	1,115,874,943	1,160,672,592	1,160,701,909	1,160,714,302	1,159,130,246
Reference genome length (bp)	1,132,614,622	1,132,614,622	1,132,614,622	1,132,614,622	1,132,614,622	1,132,614,622	1,132,614,622
GC (%)	43.50	43.50	43.50	43.50	43.51	43.51	43.50
Reference GC (%)	43.30	43.30	43.30	43.30	43.30	43.30	43.30
N50 (bp)	258,489	258,626	258,551	40,478,097	40,478,783	40,478,974	40,451,535
N90 (bp)	46,044	46,088	46,071	13,426,302	13,426,665	13,426,652	13,418,009
L50 (bp)	1,289	1,289	1,289	9	9	9	9
L90 (bp)	4,959	4,959	4,958	31	31	31	31
Number of misassemblies	3,092	3,093	3,083	2,056	2,105	2,203	2,221
Number of misassembled contigs	2,300	2,299	2,292	702	702	702	703
Misassembled contigs length (bp)	225,639,781	225,758,689	223,986,886	1,147,401,643	1,147,430,813	1,147,397,596	1,146,264,577
Number of unaligned contigs (full + partial)	176 + 2,837	176 + 2,837	175 + 2,828	160 + 873	160 + 873	164 + 875	161 + 878
Unaligned length (bp)	15,398,608	15,415,958	15,312,013	26,375,951	26,732,398	26,349,857	25,772,548
Genome fraction (%)	95.47	95.50	95.50	95.18	95.85	96.31	96.29
Duplication ratio	1.02	1.02	1.02	1.02	1.02	1.02	1.02
Number of N's per 100 kb	0.00	0.00	0.00	3,863.56	3157.46	2,703.41	2704.78
Total aligned length (bp)	1,099,763,738	1,100,104,859	1,099,674,356	1,089,366,517	1,097,164,336	1,102,778,156	1,101,784,856

2.4.3 Refinement of the draft SGH genome assembly

2.4.3.1 Contig extension of the draft assembly using SSPACE-Basic

The draft SGH genome assembly produced by DBG2OLC was further improved by extending the contigs, several iterations of polishing, scaffolding, and gap closing. Similar to the assembly stage, the performance of multiple bioinformatic tools was evaluated to produce the most improved version of the draft SGH genome assembly. Firstly, contig extension was performed using both a long-read (SSPACE-LR) and short-read (SSPACE-Basic) extension tool (Boetzer *et al*., 2011; Boetzer and Pirovano, 2014). Contig extension using SSPACE-LR with both the HiFi and CLR reads resulted in no extension of the contigs from the DBG2OLC assembly (Appendix Table 2). In comparison, SSPACE-Basic with the untrimmed and uncorrected Illumina reads (with default parameters) extended the contigs by 53.27 kb (compared to the draft DBG2OLC assembly) and merged two sets of contigs (10,516 final contigs) (Table 2.4).

The SSPACE-Basic output was further optimized to use unpaired and unmapped Illumina reads mapped using Bowtie. Mapping resulted in the production of aligned (no reads), concordantly aligned, unaligned (no reads), and concordantly unaligned reads. SSPACE-Basic was thus run with the inclusion of concordantly aligned and concordantly unaligned reads (56,822,443 unmapped reads in total) with default parameters. This resulted in an increase of N50 value by 137 bp compared to the initial draft DBG2OLC assembly and the merging of five contig pairs (Appendix Table 2). The final contig-extended draft SGH genome assembly comprised 10,513 contigs, with an improved AGH genome coverage fraction of 95.50% (Table 2.4).

2.4.3.2 Polishing of the draft assembly using Pilon and PolCa

The contig-extended SGH genome assembly was polished using Racon with the PacBio HiFi reads and Pilon with the Illumina PE reads. Polishing using Racon resulted in an increase in the number of contigs by five contigs (10,518) and a decrease in the N50 value by 3.80 kb compared to the contig extended draft assembly (Appendix Table 2). Furthermore, the genome length decreased by 17.89 Mb, while the AGH genome coverage fraction decreased by 0.27%. Polishing using Pilon resulted in a slightly better assembly compared to the Racon polished assembly with the number of contigs remaining the same and the N50 value only decreasing by 72 bp compared to the contig extended assembly. Furthermore, the first run of Pilon polishing resulted in the identification and correction of 1,060,618 errors and inaccuracies. As

such polishing using Pilon was used to improve the draft contig extended assembly over ten iterations. Over the ten iterations of Pilon, a total of 1,134,918 errors and inconsistencies were identified and corrected (Appendix Table 3). While these corrections did not affect the number of contigs (10,518) and N50 value (0.26 Mb) after the second run, the genome length decreased by 0.11 Mb and the AGH genome coverage fraction increased by 0.004% from the first to the last iteration of Pilon polishing (Appendix Table 2).

2.4.3.3 Scaffolding of the draft assembly using the AGH genome as a reference

Scaffolding of the polished draft SGH genome was conducted using the MaSurCa scaffolding tool Chromosome_Scaffolder and Ragtag using the AGH genome (GCA_009769605.1) as a reference. Chromosome_Scaffolder produced 1,808 scaffolds with an N50 value of 38.66 Mb and the length of the largest contig increasing to 0.13 Gb (Appendix Table 2). However, the genome fraction of the scaffolded assembly by Chromosome_Scaffolder decreased by 0.41% compared to the polished draft genome. Scaffolding using Ragtag outcompeted the scaffolded assembly produced by Chromosome_Scaffolder by producing a more contiguous assembly, with 155 fewer scaffolds and an N50 value of 40.48 Mb (Appendix Table 2). Furthermore, the genome length and the size of the largest contig increased by 44.78 Mb and 0.14 Gb compared to the polished draft assembly (Table 2.4). The draft assembly scaffolded by Ragtag was thus used for short- and long-read gap closing.

2.4.3.4 Gap closing of draft assembly using long- and short-read gap-closing tools

Scaffolding resulted in the introduction of gaps (Ns) (3863.56 Ns per 100 kb) within the draft assembly (Table 2.4). In order to close these gaps, long-read gap closing of the draft scaffolded SGH assembly was conducted using LR_Gapcloser, TGS-Gapcloser and the MaSurCa-integrated gap closer Samba (Xu *et al*., 2019, 2020; Zimin and Salzberg, 2022). First, LR_Gapcloser and TGS-Gapcloser were run with the independent HiFi and CLR reads to close any gaps present within the scaffolded draft assembly. LR_Gapcloser run with the independent HiFi and CLR reads resulted in an increase in the length of the largest contig by 1.62 and 3.89 kb, genome length by 12.41 and 36.38 kb and N50 value by 490 and 1,283 bp, respectively compared to the scaffolded draft assembly (Appendix Table 2). Gap closing using LR_Gapcloser also resulted in a decrease in the number of Ns by 499.82 and 691.46 Ns per 100 kb using the HiFi and CLR read datasets, respectively (Appendix Table 2). These two LR_Gapcloser assemblies outcompeted the TGS-Gapcloser assemblies run with the

independent HiFi and CLR reads, which resulted in a decrease in genome length by 1.19 and 2.35 Mb and N50 value by 67.08 and 67.49 kb, respectively compared to the scaffolded draft assembly.

As LR_Gapcloser performed the best, it was rerun with a combined HiFi and CLR dataset which resulted in a higher AGH genome coverage fraction (95.85%) compared to the runs with the independent HiFi (95.66%) and CLR (95.84%) reads (Appendix Table 2). This gap-closed assembly achieved a genome length that was 7.07 kb smaller and an N50 that was 597 bp smaller than the LR_Gapcloser assembly run with the independent CLR reads. More pertinently, it managed to resolve 14.64 more Ns per 100 kb compared to the LR_Gapcloser assembly run with the independent CLR reads.

The third gap closing tool, Samba, was also run with the combined HiFi and CLR datasets. However, while this tool did result in a slight increase in genome length (647 bp), it did not have a substantial impact on the N50 value, genome length and the number of Ns per 100 kb when compared to the scaffolded draft assembly (Appendix Table 2). As such the LR_Gapcloser output using a combination of the HiFi and CLR reads once again outcompeted the Samba output and was thus used for short-read gap closing (Table 2.4).

To further reduce the number of gaps within the long-read gap-closed draft SGH genome assembly, gap closing using the Illumina PE reads was performed using the SOAPdenovo-integrated gap closing tool GapCloser, Gap2Seq and Sealer (Luo *et al*., 2012; Paulino *et al*., 2015; Salmela and Tomescu, 2019). Of the three tools, GapCloser achieved the highest AGH genome coverage fraction (96.30%) and N50 value (40.48 Mb) (Appendix Table 2). Furthermore, Gapcloser managed to resolve 449.64 more Ns per 100 kb in the long-read gap closed draft assembly compared to Gap2Seq (0.03 Ns per 100 kb) and Sealer (6.8 Ns per 100 kb) (Appendix Table 2). To complete short-read gap closing, Sealer was used to further improve the output from GapCloser. This resulted in the resolution of 453.92 more Ns per 100 kb compared to the long-read gap closed draft assembly, 4.28 more Ns per 100 kb compared to the GapCloser-treated draft assembly alone (Appendix Table 2). The GapCloser and Sealer-combined gap-closed draft assembly further displayed an AGH genome coverage fraction of 96.31% and was thus used for the final iterations of polishing (Table 2.4).

2.4.3.5 Final polishing of the draft assembly by Pilon and PolCa

To resolve any potential errors created during gap closing using the raw PacBio and Illumina reads, a final round of polishing was conducted using five iterations of Pilon followed by five iterations of PolCa (Walker *et al*., 2014; Zimin and Salzberg, 2020). The five iterations of Pilon resulted in the identification and correction of a total of 61,437 inconsistencies and errors in the gap closed assembly (Appendix Table 4). Compared to the gap closed assembly, polishing using Pilon resulted in a slight decrease of 0.06% in the genome fraction when mapped against the AGH genome. Polishing also resulted in a decrease of 17.19 kb in the length of the largest contig and 15.09 kb in the N50 value compared to the gap closed assembly (Appendix Table 2). The size of the genome also decreased by 1.14 Mb through polishing compared to the gap closed draft assembly.

Subsequently, five iterations of PolCa run on the final Pilon polished draft assembly resulted in the identification and correction of a further 46,610 substitutions and 574,100 INDELs (Appendix Table 4). Polishing using PolCa caused no change in the number of contigs generated, but it did cause a slight increase of 0.05% in genome fraction when mapped against the AGH genome compared to polishing using Pilon (Appendix Table 2). However, the contiguity of the genome further decreased as the length of the largest contig decreased by 44.25 kb and the N50 decreased by 12.35 kb compared to the draft assembly that underwent Pilon polishing. Furthermore, the size of the genome also decreased by 0.39 Mb compared to the draft assembly that underwent Pilon polishing.

2.4.4 Comparison of the final SGH assembly and AGH assembly

The final improved and polished draft SGH genome assembly comprises 1,674 scaffolds and is 1,159,130,246 bp in size, with an N50 value of 40.45 Mb. Alignment against the AGH genome assembly, a genome fraction of 96.29% was achieved (Table 2.5). Compared to the AGH genome assembly, which comprises 41 scaffolds mapped to chromosomes, as well as 71 unplaced scaffolds, the SGH genome assembly has a greater N50 value (0.18 Mb larger) and the length of the largest contig was 1.54 Mb greater than that of the AGH assembly (Table 2.5). This suggests that the contiguity of the SGH assembly at the scaffold level is slightly better than the AGH assembly.

Table 2.5: Comparison of the SGH assembly and the AGH assembly

Assembly	Southern Ground Hornbill	Abyssinian Ground Hornbill
Number of contigs	1,653	113
Largest contig (bp)	137,990,494	136,452,362
Total length (bp)	1,159,130,246	1,132,614,622
GC (%)	43.50	43.30
N50 (bp)	40,451,535	40,274,051
N90 (bp)	13,418,009	14,657,043
L50 (bp)	9	9
L90 (bp)	31	29
Number of misassemblies	2,221	0
Number of misassembled contigs	703	0
Misassembled contigs length (bp)	1,146,264,577	0
Number of unaligned contigs	161 + 878 part	0 + 0 part
Unaligned length (bp)	25,772,548	0
Genome fraction (%)	96.29	100.00
Duplication ratio	1.02	1.00
Number of N's per 100 kbp	2,704.78	712.91
Total aligned length (bp)	1,101,784,856	1,124,539,718

Of the 1,674 scaffolds, 1,513 aligned to the draft AGH genome assembly, with a total alignment length of 1,101,784,856 bp. Alignment of the SGH against the resolved chromosomes of the AGH revealed that scaffolds with a total size of 1,131,215,034 bp aligned against these chromosomes (Appendix Table 5). This was 0.44% (4,952,660 bp) larger than the total length of all the AGH chromomeres. Comparison of individual alignments showed that a total of 33 SGH scaffolds were on average 275,497 bp larger than the AGH chromosomes. The alignment against the two sex chromosomes (W and Z) was however, on average 640,742 bp smaller. It should be noted that the SGH scaffolds had 73.67% more gaps (Ns) compared to the AGH chromosomes. The ungapped length of the AGH chromosomes was thus 1.58% (17,615,709 bp) larger than the SGH scaffolds. Furthermore, twenty-one chromosomes had a length that was on average 1,002,499 bp larger than the ungapped SGH scaffolds. A total of twenty SGH ungapped scaffolds still exhibited on average 171,838 bp larger alignment length than the ungapped chromosomes. These findings indicate that 41 SGH scaffolds display similar coverage of chromosomal arrangement as that resolved in the AGH.

The genome of the SGH is also 26.52 Mb larger than that of the AGH. Approximately 11% (2,421,303 bp) of this additional sequence can be attributed to 161 contigs which did not align with the AGH genome. These unaligned sequences have an average length of 15,039 bp, with the largest contig being 219,202 bp in length. The unaligned contigs were searched against the NCBI non-redundant nucleotide database using Blastn (Pruitt *et al*., 2007). Blastn analysis indicated that 69.57% of the contigs showed similarity to avian species mainly from the orders Accipitriformes (43.48%; mostly *Haliaeetus albicill*a – 17.39%, *Aquila chrysaetos* – 16.77% and *Accipiter gentiles* – 7.45%), Caprimulgiformes (4.35%; *Caprimulgus europaeus*), Galliformes (4.35%; *Gallus gallus*) and Coliformes (4.35%; *Phalacrocorax aristotelis*) (Figure 2.3). This suggests that the unaligned sequences may be absent from the AGH genome assembly, or potentially deleted from the genome of the latter organism.

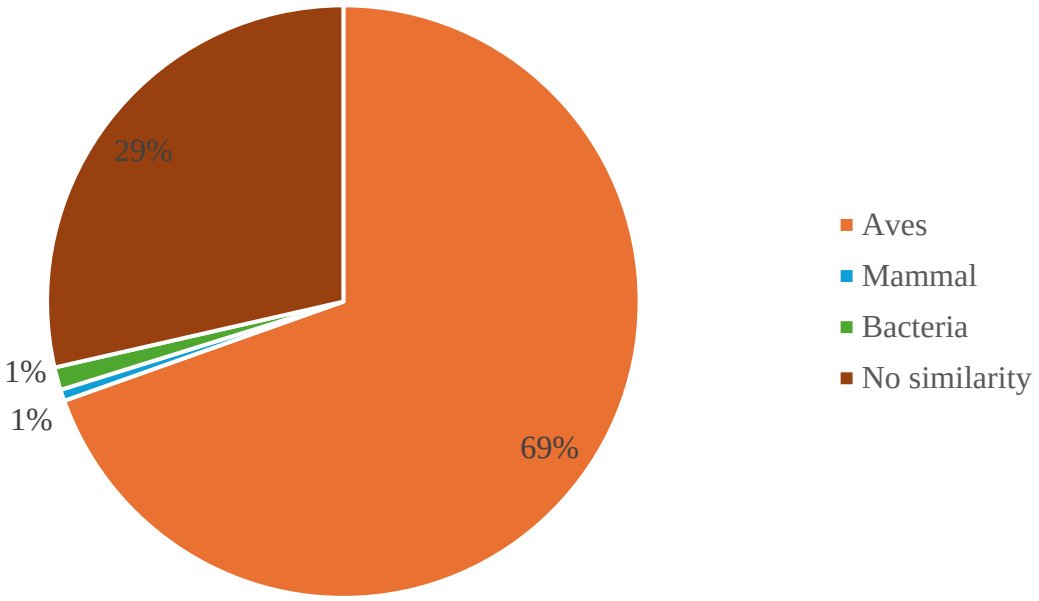


Figure 2.3: Sequence homology of the unaligned SGH contigs assembled. A total of 161 were unmapped against the reference genome of the AGH and were analysed using Blastn to identify sequence homology against other available species.

2.4.5 Genome annotations

The final SGH draft genome assembly was structurally and functionally annotated using MOSGA in an iterative process (Martin *et al*., 2021). Augustus trained with the chicken model and subsequently refined with the BUSCO output, predicted a total of 33,746 proteins with an average protein size of 399 bp. Of these proteins, 28,825 were functionally annotated by eggNOG-mapper v2 against the eggNOG 5.0 database (Cantalapiedra *et al*., 2021; Huerta-Cepas *et al*., 2019). A total of 175,151 introns were identified in the protein-coding genes, with an average of 6.83 introns per gene.

Approximately 88.60 Mb (7.64%) of the draft genome consisted of repeat sequences. Of these repetitive elements, approximately 59.09 Mb of the sequences were comprised of class I retrotransposons (mainly long interspersed nuclear elements (LINEs), 75.89% and short interspersed nuclear elements (SINEs), 7.32%) (Figure 2.4). Furthermore, approximately 3.62 Mb (3.98%) of these interspersed repeats were unclassified. No tandem repeats (satellites) were identified within the SGH draft genome.

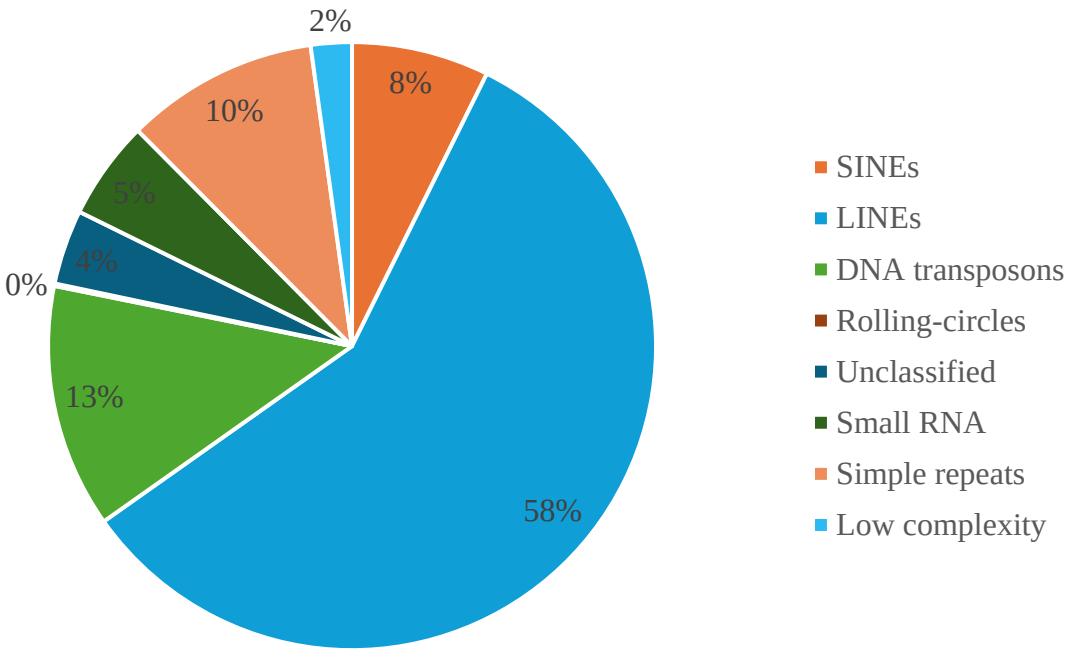
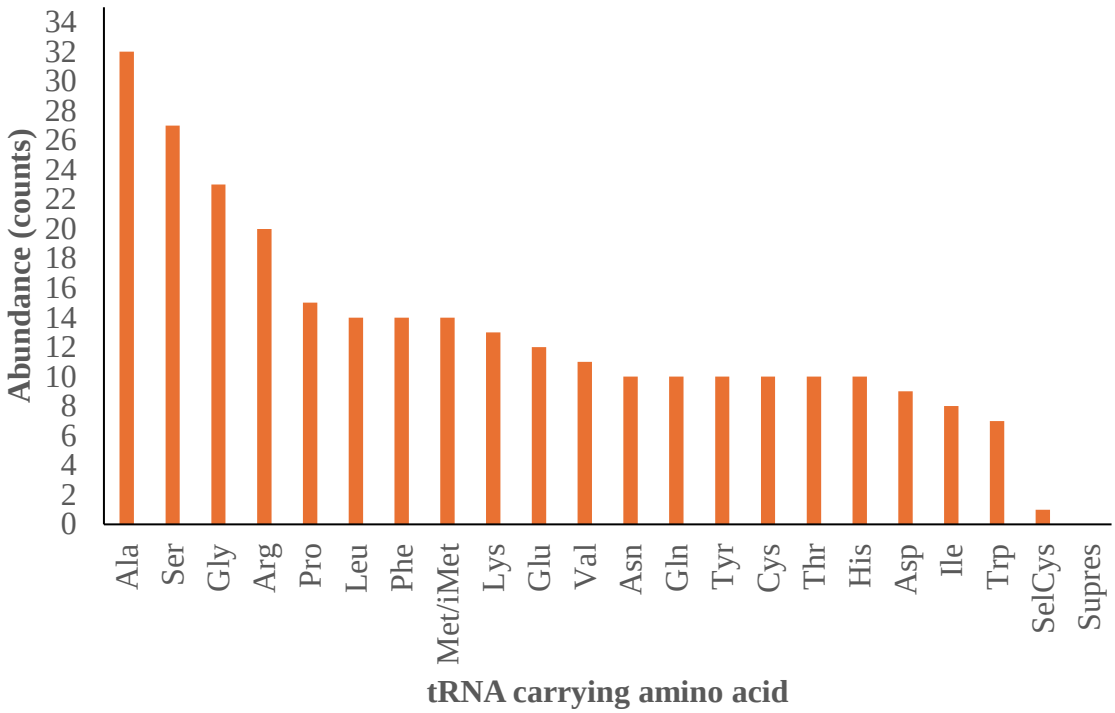


Figure 2.4: Composition of repetitive elements identified within the SGH draft genome. The annotated repeats were mainly identified from the class I retrotransposons which included the interspersed repeats (Retrotransposons – LINEs and SINEs and the DNA transposons).

A total of 280 tRNAs were predicted within the draft genome of which 162 decoded for the standard 20 amino acids. The tRNAs carrying alanine (11.43%), serine (9.64%), glycine (8.21%) and arginine (7.14%) were predominant in the SGH genome (Figure 2.5a). A total of 50 rRNAs were also predicted within the draft genome by Barrnap, which were predominated by the eukaryotic 28S (17 subtypes) and 5S (15 subtypes) rRNA subtypes (Figure 2.5b).

a)



b)

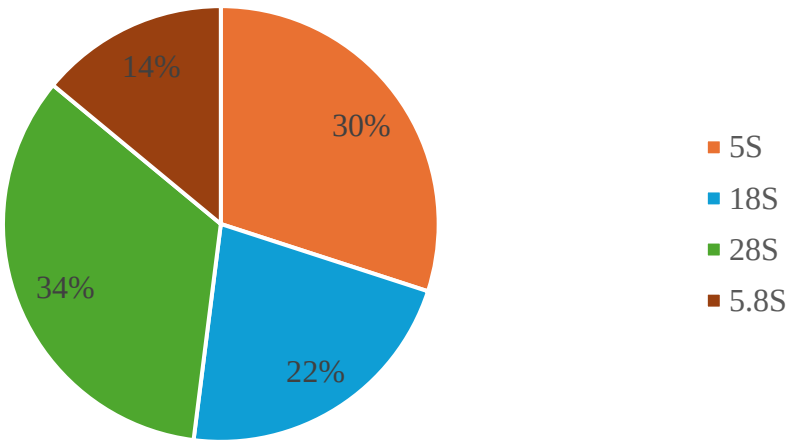


Figure 2.5: tRNA carrying amino acids and rRNA abundance within the SGH draft genome. a) tRNA-carrying amino acids are depicted in the three-letter amino acid code and the Supres label indicates the tRNA suppressor isotypes coding for stop codons. b) rRNA eukaryotic subtype composition within the SGH draft genome.

The completeness of the final SGH draft genome was assessed using BUSCO with the Avian lineage dataset composed of 8,338 genomes and 62 BUSCOs. This analysis indicated that the SGH genome assembly incorporates 7,951 (95.4%) complete BUSCOs, of which 7,910 were single-copy BUSCOs and 41 duplicated BUSCOs. A further 85 BUSCOs were identified as fragmented and 302 BUSCOs are missing.

2.5 Discussion

In this study, the complete genome of the Southern Ground Hornbill, *Bucorvus leadbeateri*, was undertaken, using both long- and short-read sequencing technologies. Subsequently, a range of assembly tools were evaluated and utilised to achieve the most accurate, contiguous, and complete draft genome of the SGH. A total of 95,695,954 raw sequenced reads were obtained using two sequencing technologies (PacBio and Illumina Novoseq) of which 94.35% were Illumina PE short reads. Assembly was undertaken with distinct short-read and long-read assemblers. The optimal independent long- and short-read assemblies achieved for the SGH both achieved a genome fraction over 89%, however, the long-read assembly did produce a more contiguous genome assembly with 106-fold less contigs and 29-fold greater N50 value compared to the short-read assembly.

Avian genomes are relatively small in size with a lower proportion of non-coding regions and repetitive elements compared to other vertebrates (Wu *et al*., 2021), however, the presence of repetitive elements can be difficult to resolve during sequencing and assembly. Short-read technologies in particular struggle to completely resolve repetitive regions due to the shorter read lengths leading to fragmented and misassembled assemblies (Haug-Baltzell *et al*., 2015; Horita *et al*., 2012; Korlach *et al*., 2017). Long-read sequencing technologies in comparison can better resolve these repetitive elements due to the longer read lengths which may span across the repeat regions (Peona *et al*., 2021). The performance of independent long-read assemblies in producing highly contiguous genomes due to their ability to resolve repeats compared to independent short-read assemblies has been noted in several studies including the Paradise crow (*Lycocorax pyrrhopterus*) (Peona *et al*., 2021), Anna's hummingbird (*Calypte anna*) (Korlach *et al*., 2017) and Steller's jay (*Cyanocitta stelleri*) (Benham *et al*., 2023).

The stand-alone assembly of the SGH genome with the long- and short-read datasets was outperformed by a hybrid *de novo* assembly approach integrating both the Illumina and PacBio reads. Similar findings of optimal assemblies being achieved via a hybrid assembly approach,

rather than independent short-read-only and long-read-only assemblies, have been reported for the Southern giant petrel (*Macronectes giganteus*) (Kim *et al*., 2021), the Barn owl (*Tyto alba* subspecies *alba*) (Ducrest *et al*., 2020) and the Northern flicker (*Colaptes auratus*) (Hruska and Manthey, 2021).

The assemblies produced in this study were evaluated based on several measures including contiguity and fragmentation, accuracy (improving errors and inconsistencies) and completeness. The contiguity of a genome relates to the size and number of contigs present within a genome in such a way that the size of the contigs is maximised and the number of contigs generated are minimised to reflect the true number and size of chromosomes in an organism (Thrash *et al*., 2020). *De novo* assembly by hybrid approach yielded a 1.12 Gb assembly comprised of 10,518 contigs. Chromosome resolution is commonly undertaken using quantitative trait loci (QTL) and genetic linkage maps. This involves the identification and mapping of genetic markers along chromosomes together with statistical analyses to determine associations between the genotypic and phenotypic data (Erickson *et al*., 2004; Fierst, 2015). However, no genetic linkage maps are available for the SGH. In such a case, further genome resolution can be undertaken by reference-guided assembly refinement with a suitable, high-quality reference genome with very little biological divergence from the assembled genome (Mikheenko *et al*., 2018; Rhie *et al*., 2021).

The genome of the sister species AGH was sequenced using PacBio Sequel I (long reads), Illumina NovaSeq (short reads), Arima Genomics Hi-C (three-dimensional organisation of genome-wide chromatin structures) and Bionano Genomics (optical maps) (Rhie *et al*., 2021). This genome has been resolved at the near-chromosome level (Rhie *et al*., 2021). The AGH genome assembly comprises 41 chromosomes, as well as 71 unplaced scaffolds. This genome was thus used for further reference-guided improvement of the *de novo* assembly of the SGH genome. This substantially improved the contiguity (6-fold decrease in number of contigs and 157-fold increase in N50), completeness and accuracy (1% increase in genome fraction) of the final draft genome. This draft genome further encompassed scaffolds that aligned against all 41 chromosomes of the AGH.

The final draft genome of the SGH is characterised by 33,746 proteins and had a BUSCO score of 95.40%. This further supports the completeness of the SGH assembly and the quality of the genome annotations indicating that the majority of the expected genes are captured in the assembly. In addition to protein-coding genes, the genome also encompassed 280 tRNAs and

50 rRNAs which are essential non-coding RNAs that play crucial roles in protein synthesis (Catalanotto *et al*., 2023; Ottenburghs *et al*., 2021). The draft genome was further composed of 7.64% repetitive elements, a proportion within the expected range for avian species, and indicative of potential regulatory roles these elements might play in gene expression (Zhang *et al*., 2014). Overall, these results indicate the high quality of the SGH genome assembly and annotations and provide an overview of the genetic blueprint of this species.

While a well-resolved genome sequence and assembly of the SGH has been achieved at the scaffold level, there is potential for further refinement. This assembly could be improved to a chromosome-level assembly, similar to that of the AGH. This process would involve connecting scaffolds into larger blocks corresponding to chromosomes. This can be achieved using linked reads (such as those provided by 10X Genomics), optical maps (like Bionano optical maps and Arima Genomics) and Hi-C data (from Dovetail Genomics and Phase Genomics Hi-C) (Rhie *et al*., 2021). Additionally, the development of unique primers targeting remaining gap regions or difficult-to-resolve repeat-rich regions and subsequent sequencing could be conducted to further improve the final chromosome-level assembly (Garber *et al*., 2009; Zhang *et al*., 2012). The resulting assembly would provide a more complete picture of the SGH's genome structure. However, it is important to note that achieving chromosome-level assemblies is not always feasible due to the additional costs and computational resources required for this process (Etherington *et al*., 2020). Despite this, the scaffold-level assembly of the SGH genome still provides a crucial platform for the study of the genetic make-up of the SGH and to identify the genetic pathways and mechanisms driving its adaptation and survival in its environment. Furthermore, it serves as a valuable reference genome that can be used to resolve more SGH genomes in the future, providing a solid foundation for population genomic studies. These studies can help better understand the genetic diversity, population structure, phylogeography and demography of SGH populations. In turn, these insights can contribute towards conservation efforts for this species by providing a deeper understanding of its genetic diversity and adaptation.

2.6 Conclusion

We present the first sequenced, *de novo* assembled and annotated genome of the SGH achieved using a combination of Illumina short-reads and PacBio long-reads, two sequencing technologies commonly used in assembling avian genomes. The combination of both long- and short-reads with a hybrid *de novo* assembly approach was superior in resolving the whole genome compared to stand-alone assemblers. Despite achieving a genome coverage that was relatively low, a highly contiguous and accurate genome was achieved. This was made possible by parameter optimisation and manual curation of state-of the art assemblers and assembly improvement tools. Using this approach, parameters and tools best suited for the data generated were identified and utilised. Overall, the draft SGH genome assembly provides a pivotal foundation for resolving the genetic blueprint of this species to identify genetic pathways and mechanisms used in their survival. This will aid current conservation efforts and ultimately preserve the history of the SGH in the possible event of extinction.

2.7 References

Aird, D., Ross, M. G., Chen, W. S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology*, *12*(2), R18. https://doi.org/10.1186/gb-2011-12-2-r18

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z. B., Schatz, M. C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology*, *23*(1), 258. https://doi.org/10.1186/s13059-022-02823-7

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., and Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), Article 1. https://doi.org/10.1186/s13059-019-1829-6

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., and Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, *21*(1), 30. https://doi.org/10.1186/s13059-020-1935-5

Andrew, S. (2015). *FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]* [Computer software]. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Angel, V. D. D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O. V., Amselem, J., Bouri, L., Bocs, S., Klopp, C., Gibrat, J. F., Vlasova, A., Leskosek, B. L., Soler, L., Binzer-Panchal, M., and Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research*, *7*. https://doi.org/10.12688/f1000research.13598.1

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Athanasopoulou, K., Boti, M. A., Adamopoulos, P. G., Skourou, P. C., and Scorilas, A. (2022). Third-Generation Sequencing: The Spearhead towards the Radical Transformation of Modern Genomics. *Life*, *12*(1), Article 1. https://doi.org/10.3390/life12010030

B10K. (2024). *B10K Database*. https://b10k.genomics.cn/progress.html

Benham, P. M., Cicero, C., DeRaad, D. A., McCormack, J. E., Wayne, R. K., Escalona, M., Beraut, E., Marimuthu, M. P. A., Nguyen, O., Nachman, M. W., and Bowie, R. C. K. (2023). A highly contiguous reference genome for the Steller's jay (Cyanocitta stelleri). *Journal of Heredity*, *114*(5), 549–560. https://doi.org/10.1093/jhered/esad042

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), 623–630. https://doi.org/10.1038/nbt.3238

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics*, *27*(4), 578–579. https://doi.org/10.1093/bioinformatics/btq683

Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics*, *15*(1), 211. https://doi.org/10.1186/1471-2105-15-211

Burt, D. W. (2005). Chicken genome: Current status and future opportunities. *Genome Research*, *15*(12), 1692–1698. https://doi.org/10.1101/gr.4141805

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), 5825–5829. https://doi.org/10.1093/molbev/msab293

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for Glycogenomics. *Nucleic Acids Research*, *37*(Database issue), D233–D238. https://doi.org/10.1093/nar/gkn663

Catalanotto, C., Barbato, C., Cogoni, C., and Benelli, D. (2023). The RNA-Binding Function of Ribosomal Proteins and Ribosome Biogenesis Factors in Human Health and Disease. *Biomedicines*, *11*(11), 2969. https://doi.org/10.3390/biomedicines11112969

Chaisson, M. J., and Tesler, G. (2012). Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): Application and theory. *BMC Bioinformatics*, *13*(1), 238. https://doi.org/10.1186/1471-2105-13-238

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. (2021). tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096. https://doi.org/10.1093/nar/gkab688

Chan, P. P., and Lowe, T. M. (2019). tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods in Molecular Biology (Clifton, N.J.)*, *1962*, 1–14. https://doi.org/10.1007/978-1-4939-9173-0_1

Cheng, Y., and Burt, D. W. (2018). Chicken genomics. *The International Journal of Developmental Biology*, *62*(1-2–3), Article 1-2–3. https://doi.org/10.1387/ijdb.170276yc

Chikhi, R., and Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31–37. https://doi.org/10.1093/bioinformatics/btt310

Delmore, K. E., and Liedvogel, M. (2020). Avian Population Genomics Taking Off: Latest Findings and Future Prospects. In J. Y. Dutheil (Ed.), *Statistical Population Genomics* (eds, Vol. 2090, pp. 413–433). Springer US. https://doi.org/10.1007/978-1-0716-0199-0_17

Deng, Z., and Delwart, E. (2021). ContigExtender: A new approach to improving de novo sequence assembly for viral metagenomics data. *BMC Bioinformatics*, *22*(1), 119. https://doi.org/10.1186/s12859-021-04038-2

Ducrest, A. L., Neuenschwander, S., Schmid-Siegert, E., Pagni, M., Train, C., Dylus, D., Nevers, Y., Warwick Vesztrocy, A., San-Jose, L. M., Dupasquier, M., Dessimoz, C., Xenarios, I., Roulin, A., and Goudet, J. (2020). New genome assembly of the barn owl (Tyto alba alba). *Ecology and Evolution*, *10*(5), 2284–2298. https://doi.org/10.1002/ece3.5991

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Erickson, D. L., Fenster, C. B., Stenøien, H. K., and Price, D. (2004). Quantitative trait locus analyses and the study of evolutionary process. *Molecular Ecology*, *13*(9), 2505–2522. https://doi.org/10.1111/j.1365-294X.2004.02254.x

Etherington, G., Heavens, D., Baker, D., Lister, A., McNelly, R., Garcia, G., Clavijo, B., Macaulay, I., Haerty, W., and Di Palma, F. (2020). Sequencing smart: De novo sequencing and assembly approaches for a non-model mammal. *GigaScience*, *9*. https://doi.org/10.1093/gigascience/giaa045

Fierst, J. L. (2015). Using linkage maps to correct and scaffold de novo genome assemblies: Methods, challenges, and computational tools. *Frontiers in Genetics*, *6*. https://doi.org/10.3389/fgene.2015.00220

Garber, M., Zody, M. C., Arachchi, H. M., Berlin, A., Gnerre, S., Green, L. M., Lennon, N., *and* Nusbaum, C. (2009). Closing gaps in the human genome using sequencing by synthesis. *Genome Biology*, *10*(6), R60. https://doi.org/10.1186/gb-2009-10-6-r60

Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., … Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1), iyad031. https://doi.org/10.1093/genetics/iyad031

Gupta, N., and Verma, V. K. (2019). Next-Generation Sequencing and Its Application: Empowering in Public Health Beyond Reality. *Microbial Technology for the Welfare of Society*, *17*, 313–341. https://doi.org/10.1007/978-981-13-8844-6_15

Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Hannon Lab. (n.d.). *FASTX-Toolkit* [Computer software]. http://hannonlab.cshl.edu/fastx_toolkit/

Haug-Baltzell, A., Jarvis, E. D., McCarthy, F. M., and Lyons, E. (2015). Identification of dopamine receptors across the extant avian family tree and analysis with other clades uncovers a polyploid expansion among vertebrates. *Frontiers in Neuroscience*, *9*. https://www.frontiersin.org/articles/10.3389/fnins.2015.00361

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., and Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 61-passim. https://doi.org/10.2144/000114133

Heather, J. M., and Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Horita, H., Kobayashi, M., Liu, W., Oka, K., Jarvis, E. D., and Wada, K. (2012). Specialized Motor-Driven dusp1 Expression in the Song Systems of Multiple Lineages of Vocal Learning Birds. *PLOS ONE*, *7*(8), e42173. https://doi.org/10.1371/journal.pone.0042173

Hruska, J. P., and Manthey, J. D. (2021). De novo assembly of a chromosome-scale reference genome for the northern flicker Colaptes auratus. *G3 Genes|Genomes|Genetics*, *11*(1), jkaa026. https://doi.org/10.1093/g3journal/jkaa026

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., von Mering, C., and Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., Jahesh, G., Khan, H., Coombe, L., Warren, R. L., and Birol, I. (2017). ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. *Genome Research*, *27*(5), 768–777. https://doi.org/10.1101/gr.214346.116

Jung, H., Ventura, T., Chung, J. S., Kim, W. J., Nam, B. H., Kong, H. J., Kim, Y. O., Jeon, M. S., and Eyun, S. (2020). Twelve quick steps for genome assembly and annotation in the

classroom. *PLOS Computational Biology*, *16*(11), e1008325. https://doi.org/10.1371/journal.pcbi.1008325

Kammonen, J. I., Smolander, O. P., Paulin, L., Pereira, P. A. B., Laine, P., Koskinen, P., Jernvall, J., and Auvinen, P. (2019). gapFinisher: A reliable gap filling pipeline for SSPACE-LongRead scaffolder output. *PLoS ONE*, *14*(9), e0216885. https://doi.org/10.1371/journal.pone.0216885

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*(suppl_1), D480–D484. https://doi.org/10.1093/nar/gkm882

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kchouk, M., Gibrat, J. F., and Elloumi, M. (2017). Generations of Sequencing Technologies: From First to Next Generation. *Biology and Medicine*, *09*. https://doi.org/10.4172/0974-8369.1000395

Kim, S. H., Lee, S. J., Jo, E., Kim, J., Kim, J. U., Kim, J. H., Park, H., and Chi, Y. M. (2021). Genome of the Southern Giant Petrel Assembled Using Third-Generation DNA Sequencing and Linked Reads Reveals Evolutionary Traits of Southern Avian. *Animals*, *11*(7), Article 7. https://doi.org/10.3390/ani11072046

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), 722–736. https://doi.org/10.1101/gr.215087.116

Korlach, J., Gedman, G., Kingan, S. B., Chin, C. S., Howard, J. T., Audet, J. N., Cantin, L., and Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*, *6*(10), gix085. https://doi.org/10.1093/gigascience/gix085

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25. https://doi.org/10.1186/gb-2009-10-3-r25

Li, H. (2014). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. 0 Bytes. https://doi.org/10.6084/M9.FIGSHARE.963153.V1

Li, H. (2016). Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. *Bioinformatics*, *32*(14), 2103–2110. https://doi.org/10.1093/bioinformatics/btw152

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., Gan, J., Li, N., Hu, X., Liu, B., Yang, B., and Fan, W. (2012). Comparison of the two major classes of assembly algorithms: Overlap–layout–consensus and de-bruijn-graph. *Briefings in Functional Genomics*, *11*(1), 25–37. https://doi.org/10.1093/bfgp/elr035

Lin, B., Hui, J., and Mao, H. (2021). Nanopore Technology and Its Applications in Gene Sequencing. *Biosensors*, *11*(7), Article 7. https://doi.org/10.3390/bios11070214

Logsdon, G. A., Vollger, M. R., and Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews. Genetics*, *21*(10), 597–614. https://doi.org/10.1038/s41576-020-0236-x

Luo, J., Wei, Y., Lyu, M., Wu, Z., Liu, X., Luo, H., and Yan, C. (2021). A comprehensive review of scaffolding methods in genome assembly. *Briefings in Bioinformatics*, *22*(5), bbab033. https://doi.org/10.1093/bib/bbab033

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., Tang, J., Wu, G., Zhang, H., Shi, Y., Liu, Y., Yu, C., Wang, B., Lu, Y., Han, C., … Wang, J.

(2012). SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *1*, 18. https://doi.org/10.1186/2047-217X-1-18

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., and Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Martin, R., Hackl, T., Hattab, G., Fischer, M. G., and Heider, D. (2021). MOSGA: Modular Open-Source Genome Annotator. *Bioinformatics*, *36*(22–23), 5514–5515. https://doi.org/10.1093/bioinformatics/btaa1003

Mikheenko, A., Prjibelski, A., Saveliev, V., Antipov, D., and Gurevich, A. (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics*, *34*(13), i142–i150. https://doi.org/10.1093/bioinformatics/bty266

Ottenburghs, J., Geng, K., Suh, A., and Kutter, C. (2021). Genome Size Reduction and Transposon Activity Impact tRNA Gene Diversity While Ensuring Translational Stability in Birds. *Genome Biology and Evolution*, *13*(4), evab016. https://doi.org/10.1093/gbe/evab016

Pareek, C. S., Smoczynski, R., and Tretyn, A. (2011). Sequencing technologies and genome sequencing. *Journal of Applied Genetics*, *52*(4), 413–435. https://doi.org/10.1007/s13353-011-0057-x

Paulino, D., Warren, R. L., Vandervalk, B. P., Raymond, A., Jackman, S. D., and Birol, I. (2015). Sealer: A scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics*, *16*(1), 230. https://doi.org/10.1186/s12859-015-0663-4

Peona, V., Blom, M. P. K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K. A., Zhou, Q., Irestedt, M., and Suh, A. (2021). Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Molecular Ecology Resources*, *21*(1), 263–286. https://doi.org/10.1111/1755-0998.13252

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *35*(Database issue), D61–D65. https://doi.org/10.1093/nar/gkl842

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D. (2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), Article 7856. https://doi.org/10.1038/s41586-021-03451-0

Rhoads, A., and Au, K. F. (2015). PacBio Sequencing and Its Applications. *Genomics, Proteomics and Bioinformatics*, *13*(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nature Methods*, *17*(2), Article 2. https://doi.org/10.1038/s41592-019-0669-3

Salmela, L., and Tomescu, A. I. (2019). Safely Filling Gaps with Partial Solutions Common to All Solutions. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *16*(2), 617–626. https://doi.org/10.1109/TCBB.2017.2785831

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), 293. https://doi.org/10.1186/s12864-020-6707-9

Schatz, M. C., Delcher, A. L., and Salzberg, S. L. (2010). Assembly of large genomes using second-generation sequencing. *Genome Research*, *20*(9), 1165–1173. https://doi.org/10.1101/gr.101360.109

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, *95*(11), 5857–5864. https://doi.org/10.1073/pnas.95.11.5857

Seemann, T. (2013). *Barrnap* [Perl]. https://github.com/tseemann/barrnap (Original work published 2013)

Shieh, Y. K., Liu, S. C., Lu, C. L., Shieh, Y. K., Liu, S. C., and Lu, C. L. (2020). Scaffolding Contigs Using Multiple Reference Genomes. In *Computational Biology and Chemistry*. IntechOpen. https://doi.org/10.5772/intechopen.93456

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), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., and Birol, İ. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123. https://doi.org/10.1101/gr.089532.108

Smit, A. F. A., Hubley, R., and Green, P. (1996). *RepeatMasker*. RepeatMasker. http://www.repeatmasker.org

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*(3), 405–420. https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(suppl_2), W435–W439. https://doi.org/10.1093/nar/gkl200

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics*, *25*(1). https://doi.org/10.1002/0471250953.bi0410s25

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, *28*(1), 33–36.

Thrash, A., Hoffmann, F., and Perkins, A. (2020). Toward a more holistic method of genome assembly assessment. *BMC Bioinformatics*, *21*(4), 249. https://doi.org/10.1186/s12859-020-3382-4

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), 737–746. https://doi.org/10.1101/gr.214270.116

Verma, M., Kulshrestha, S., and Puri, A. (2017). Genome Sequencing. In J. M. Keith (Ed.), *Bioinformatics: Volume I: Data, Sequence Analysis, and Evolution* (Vol. 1525, pp. 3–33). Springer. https://doi.org/10.1007/978-1-4939-6622-6_1

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., and Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), Article 11. https://doi.org/10.1371/journal.pone.0112963

Walve, R., Salmela, L., and Mäkinen, V. (2017). Variant genotyping with gap filling. *PLOS ONE*, *12*(9), e0184608. https://doi.org/10.1371/journal.pone.0184608

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., and Zdobnov, E. M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X. J., Buck, D., and Au, K. F. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research*, *6*, 100. https://doi.org/10.12688/f1000research.10571.2

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., Töpfer, A., Alonge, M., Mahmoud, M., Qian, Y., Chin, C. S., Phillippy, A. M., Schatz, M. C., Myers, G., DePristo, M. A., … Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology*, *37*(10), 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Wheeler, T. J., and Eddy, S. R. (2013). nhmmer: DNA homology search with profile HMMs. *Bioinformatics*, *29*(19), 2487–2489. https://doi.org/10.1093/bioinformatics/btt403

Wu, L., Jiao, X., Zhang, D., Cheng, Y., Song, G., Qu, Y., and Lei, F. (2021). Comparative Genomics and Evolution of Avian Specialized Traits. *Current Genomics*, *22*(7), 496–511. https://doi.org/10.2174/1389202923666211227143952

Xu, G. C., Xu, T. J., Zhu, R., Zhang, Y., Li, S. Q., Wang, H. W., and Li, J. T. (2019). LR_Gapcloser: A tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience*, *8*(1), giy157. https://doi.org/10.1093/gigascience/giy157

Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B. A., Fan, G., Liu, X., Xu, X., Deng, L., and Zhang, Y. (2020). TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience*, *9*(9), giaa094. https://doi.org/10.1093/gigascience/giaa094

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. (Sam). (2016). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep31900

Ye, C., and Ma, Z. S. (2016). Sparc: A sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ*, *4*, e2016. https://doi.org/10.7717/peerj.2016

Ye, C., Ma, Z. S., Cannon, C. H., Pop, M., and Yu, D. W. (2012). Exploiting sparseness in de novo genome assembly. *BMC Bioinformatics*, *13*(6), S1. https://doi.org/10.1186/1471-2105-13-S6-S1

Zhang, X., Davenport, K. W., Gu, W., Daligault, H. E., Munk, A. C., Tashima, H., Reitenga, K., Green, L. D., and Han, C. S. (2012). Improving Genome Assemblies by Sequencing PCR Products with PacBio. *BioTechniques*, *53*(1), 61–62. https://doi.org/10.2144/0000113891

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., … Wang, J. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), Article 6215. https://doi.org/10.1126/science.1251385

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., and Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

Zimin, A. V., and Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology*, *16*(6), e1007981. https://doi.org/10.1371/journal.pcbi.1007981

Zimin, A. V., and Salzberg, S. L. (2022). The SAMBA tool uses long reads to improve the contiguity of genome assemblies. *PLOS Computational Biology*, *18*(2), e1009860. https://doi.org/10.1371/journal.pcbi.1009860

Chapter three – Comparative genomic analysis of the Southern Ground Hornbill

3.1 Abstract

Comparative genomics has played a major role in our understanding of avian species and their evolution from other vertebrates at a genome-wide scale. However, these studies have primarily focused on major avian orders, with smaller orders such as the Bucerotiformes receiving little attention. To gain insight into the unique biological traits, adaptations and evolutionary history of these iconic birds, comparative analysis of the available genome sequences of members of the genera *Bucorvus* (Southern Ground Hornbill and Abyssinian Ground Hornbill) and *Buceros* (Rhinoceros hornbill and the Great hornbill) was performed. The *Bucorvus* genomes were on average 60.52 Mb larger than those of the *Buceros* species and harbour higher numbers of protein-coding genes, tRNAs, rRNAs and repeat content, suggesting a greater genome complexity for members of the former genus. Functional analysis of the accessory genome highlighted a more complex set of proteins associated with biological pathways including lipid and energy metabolism, thermogenesis, protein digestion and signal transduction in the *Bucorvus* species compared to the *Buceros* species. This could have contributed to the adaptation of the *Bucorvus* species to the African savannah biome following divergence from other hornbill species. Ultimately, understanding the unique biological traits, adaptations, and evolutionary history of *Bucorvus* and *Buceros* species can assist in informing conservation strategies for these vulnerable birds tailored to their specific needs in their respective environments.

3.2 Introduction

Comparative genomics encompasses a range of approaches that allow for the comprehensive comparison of two or more genomes of species to identify differences and similarities between them (Wei *et al*., 2002). Since the sequencing of the first bird genome (chicken) in 2004 (Hillier *et al*., 2004), these approaches have played a major role in avian genomics. With ever-increasing numbers of avian genomes being sequenced, comparative genomic methodologies have and will continue to provide unique insights into the biology, ecology and evolution of birds (Bravo *et al*., 2021; Feng *et al*., 2020; Stiller and Zhang, 2019; Wu *et al*., 2021; Zhang *et al*., 2014b).

Wide-scale comparative genomic analysis of avian species has been conducted at several levels including the comparison of avian species with other vertebrates, which has helped elucidate avian-specific patterns of genome evolution and traits (Bravo *et al*., 2021). Furthermore, a

comparison between major avian orders has highlighted lineage-specific traits (Bravo *et al*., 2021; Wu *et al*., 2021). While such studies have substantially improved our understanding of avian evolution as a whole (Bravo *et al*., 2021), there is still a major gap in knowledge as comparative genomic analysis of avian species has mainly focused on major avian orders, using representative species (Jarvis *et al*., 2014; Oliveros *et al*., 2019; Prum *et al*., 2015; Zhang *et al*., 2014a). Comparative genomic analysis at a genus and species level of smaller avian orders is only now coming to the fore (Cheng *et al*., 2020; Eliason *et al*., 2023; Harvey *et al*., 2020; Vianna *et al*., 2020). The order Bucerotiformes is one such avian order whose evolutionary history and molecular biology has received little attention, with molecular studies based on short DNA fragments rather than the entire genome (Delport *et al*., 2002; Gonzalez *et al*., 2013; Lan *et al*., 2019; Sammler *et al*., 2011).

To gain insight into the unique biological traits, adaptations and evolutionary history of members of the order Bucerotiformes, we performed comparative genomic analysis. This analysis used the draft genome assembly of the Southern Ground Hornbill (SGH) from Chapter 2 and compared it against the genome sequences of three other hornbill species which are publicly available.

3.3 Methods and materials

3.3.1 Genome sequences, structural and function annotation

The genome assemblies of three hornbill species are publicly available in the NCBI Assembly Database (www.ncbi.nlm.nih.gov/assembly). This includes two representatives of the family *Bucerotidae* and genus *Buceros*, namely *Buceros bicornis* (Great hornbill - GH; GCA_027563805.1; ASM2756380v1; Iridian Genomes, 2023) and *Buceros rhinoceros* subsp. *silvestris* (Rhinoceros hornbill – RH; GCF_000710305.1; ASM71030v1; Jarvis *et al*., 2014). *Bucorvus abyssinicus* (Abyssinian Ground Hornbill – AGH; GCA_009769605.1; bBucAby1.pri; Rhie *et al*., 2021) forms the second representative of the family *Bucorvidae* and genus *Bucorvus*. The reference genome sequence of *Gallus gallus* (chicken; GCF_016699485.2; bGalGal1; Vertebrate Genomes Project, 2023) was included as an outgroup for phylogenetic analysis.

All comparator genome sequences were structurally annotated in the same manner as the SGH draft genome (**Chapter 2 section 2.3.6**) using the annotation tool MOSGA (Martin *et al*., 2021) in an iterative process. In brief, repetitive elements were identified and classified using

RepeatMasker v 4.1.1 (Smit *et al*., 1996) soft-masked using the chicken genome as a reference. The soft-masked genome was used for *ab initio* gene prediction by Augustus trained with the chicken genome (Stanke *et al*., 2006). The BUSCO output was further used to retrain Augustus to improve the gene models and protein prediction (Simão *et al*., 2015). tRNA and rRNA were predicted using tRNAscan-SE v 2.0 (Chan *et al*., 2021) with default parameters and Barrnap v 0.9 (Seemann and Booth, 2013), set with a minimum E-value of 10e-6, to identify conserved eukaryotic 5S, 5.8S, 28S and 18S rRNA.

The SGH, AGH, RH and GH genomes were functionally annotated using the eggNOG-mapper v 2 (Cantalapiedra *et al*., 2021) implementation in MOSGA to assign biochemical functions to the identified genes (Martin *et al*., 2021) as per **Chapter 2 section 2.3.6**. This algorithm annotates the predicted genes using Conserved Orthologous Groups (COGs) categories (Tatusov *et al*., 2000), Gene ontology (GO) terms (Ashburner *et al*., 2000; Gene Ontology Consortium *et al*., 2023), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa and Goto, 2000), Simple Modular Architecture Research Tool (SMART)/PFAM domains (Schultz *et al*., 1998; Sonnhammer *et al*., 1997), Carbohydrate-active enzymes (CAZy) (Cantarel *et al*., 2009) and KEGG modules (Kanehisa *et al*., 2008).

Genome completeness and the quality of the predicted gene models were assessed using BUSCO v. 5.6.1 (Simão *et al*., 2015) with default parameters using the Aves lineage dataset aves_odb10 (number of genes: 8,338).

3.3.2 Identification of orthologues and pan-genome analysis

The protein datasets derived by structural annotation (**section 3.3.1**) were used to identify orthologous proteins conserved between the different hornbills with OrthoFinder v 2.5.5 (Emms and Kelly, 2019). Orthologues are structured into orthogroups, which encompass both orthologues (genes that diverged via speciation which conserve their biological function) and paralogues (genes that diverged via gene duplication which may evolve new functions) (Emms and Kelly, 2015; Gabaldón and Koonin, 2013). Comparative analysis of these orthogroups allows for the identification of overlapping clusters across a set of species. These overlaps permit the establishment of relationships between three key elements, the structure of proteins, the function of genes and the taxonomic classification of the species, providing insight into the evolution of these elements over time (Gabaldón and Koonin, 2013; Koonin, 2005).

In brief, an all-vs-all Blastp search, with an e-value threshold of 10^{-3}, was performed with the protein datasets from each genome to identify proteins with sequence similarity above the threshold (Johnson *et al*., 2008). The Blast hits were then normalised against protein length and phylogenetic distance to reduce length and distance dependency (Emms and Kelly, 2015). The reciprocal best length-normalised hit (RBNH) value was used to set a lower limit threshold for the sequence similarity scores against which the hits were sorted. Hits with a ≥ RBNH score were retained and used to construct an orthologous graph with the normalised hits as weights. This graph was then used to cluster the proteins into orthogroups using the Markov Cluster (MCL) algorithm (Emms and Kelly, 2015).

The OrthoFinder output for all compared hornbill species was used to analyse the pan-genome of these species. The pan-genome represents the sum of all proteins shared by species (core genome), proteins shared by some species but not all species (accessory genome) and the organism-specific proteins (unique genome) within the comparator dataset (Tettelin *et al*., 2008). COG annotations from the eggNOG functional annotation output were retrieved for the core, accessory and unique protein datasets and the proportion of each was calculated.

Since the primary focus of this chapter was to identify differences between the hornbill genomes at a genus level, the accessory genome between the *Bucorvus* and *Buceros* species was analysed in-depth. The KEGG Orthology (KO) terms characterised in the eggNOG annotations were utilised to identify differences in protein function between these birds. These KO terms were used to reconstruct the KEGG pathways with Reconstruct from KEGG mapper v 5 tools (Kanehisa *et al*., 2022; Kanehisa and Sato, 2020). KEGG pathways that exhibited a five-fold difference in the abundance of proteins in either the genus *Bucorvus* or *Buceros* compared to the alternate genus were used for functional analysis.

3.3.3 Gene duplication events

Paralogous proteins, those with multiple copies in one/or more genomes, were identified from the OrthoFinder output. The functional implications of those orthogroups with five-fold greater protein copies in either the genus *Bucorvus* or *Buceros* compared to the alternate genus were further analysed using the COG and KO term (KEGG pathways) annotations derived from eggNOG.

3.3.4 Core genome phylogeny

Single-copy orthologous proteins (SCOs – conserved proteins present in only one copy in all compared taxa) were extracted from the protein datasets. Each SCO dataset was individually aligned using the m-Coffee implementation of T-Coffee v. 13.46.0.919e8c6b (Di Tommaso *et al*., 2011; Wallace *et al*., 2006). The SCO alignments were concatenated using an in-house Perl script. Poorly aligned blocks and large gaps were removed from the concatenated alignment using Gblocks v. 0.91b (Talavera and Castresana, 2007).

The curated concatenated SCO alignment was used to construct a maximum-likelihood (ML) phylogeny with IQ-Tree v 2.0.7 (Minh *et al*., 2020). In brief, IQ-Tree first performed model selection with ModelFinder to identify the evolutionary model that best fitted the dataset provided using a combination of substitution and rate-heterogeneity-across-sites models (Kalyaanamoorthy *et al*., 2017). A ML phylogenetic tree was then constructed and optimised using the hill climbing nearest neighbour interchange (NNI) approach (Nguyen *et al*., 2015). In this approach, the internal branches of the species tree were swapped, and the likelihood of the tree was computed, a process that was repeated to explore different tree topologies until the best tree model was identified. Lastly, the reliability of the branches was evaluated with bootstrap support, using UFBoot (Hoang *et al*., 2018) with 1,000 replicates.

3.3.5 Comparative analysis of hornbill mitochondrial genomes

The complete mitochondrial genome (mtDNA) sequences of the AGH (GenBank: MN356327.1; Feng *et al*., 2020), GH (GenBank: MF325011.1; Chen *et al*., 2018) and RH (GenBank: MG596878.1; Lan *et al*., 2019) are publicly available on the NCBI Nucleotide Database (www.ncbi.gov/nuccore). Furthermore, the chicken mtDNA (NCBI Reference Sequence: NC_053523.1; Warren *et al*., 2023) was used as an outgroup for comparison. The mitochondrial genome of the SGH (NCBI Reference Sequence: NC_015199.1; Pacheco *et al*., 2011) is also available on the NCBI database. Blastn was performed with this sequence against the raw SGH short (Illumina) and long (Pacific BioSciences – PacBio) reads from **Chapter 2 section 2.3.3**, as well as the *de novo* short-read and hybrid assembly, using localised Blastn. The mtDNA genome of our SGH was found to be 100% identical to that uploaded on the NCBI database, which was subsequently used.

The mtDNAs of the four hornbill species and the chicken were aligned using the MAFFT v 7 server (Katoh *et al*., 2019) with default parameters. As the mitochondrial genomes are all circular, the alignments were shifted to all commence in agreement with the first nucleotide

position in the SGH mitochondrial genome. Poorly aligned and gap regions were removed from the mtDNAs using Gblocks v. 0.91b (Talavera and Castresana, 2007). The resultant alignment was used to construct a ML phylogenetic tree using PhyML v 3.0 (Guindon *et al*., 2010). In brief, PhyLM first performed automatic model selection with Smart model selection (SMS) (Lefort *et al*., 2017) using the Bayesian information criterion (Neath and Cavanaugh, 2012). A ML phylogenetic tree was then constructed as per **section 3.3.4**. Lastly, the reliability of the branches was evaluated with bootstrap support (Guindon *et al*., 2010) with 1,000 replicates.

The mtDNAs of the four hornbills and the chicken were annotated using MITOS2 via the Galaxy v 2.1.9+galaxy0 server (Al Arab *et al*., 2017; Donath *et al*., 2019) with default parameters. The mtDNA FASTA files and GenBank flat file annotations were used to generate GenBank files using the *seqret* command line tool from EMBOSS (EMBOSS, 2002). A Blast atlas was generated using the SGH mitochondrial genome as reference and Blastn was performed against the other mitochondrial genomes in 50, 100, 250 and 500 bp windows with default parameters. The Blast results were then used to map regions of homology using CGview (Stothard and Wishart, 2005).

3.4 Results

3.4.1 Hornbill genome metrices

Comparison of the genome assemblies of the four hornbill species indicated that the *Bucorvus* species have substantially larger genomes (average 60.52 Mb larger – 5.58 % - than the latter taxa), with a slightly higher G+C content (0.49 % higher on average) compared to their *Buceros* counterparts (Table 3.1). By proxy, the *Bucorvus* genomes also coded for an average of 7,853.5 more proteins than the two *Buceros* species. Of note, while the genomes of the *Bucorvus* species are 5.58% larger than the *Buceros* species, they code for 30.31% more proteins than the latter taxa. The protein-coding genes of the *Bucorvus* species also included more introns (173,899 total introns, 6.87 introns/gene) than their *Buceros* counterparts (145,418 total introns, 7.00 introns/gene).

Table 3.1: Structural annotation statistics of the hornbill genomes

Genome matrices	Southern Ground Hornbill	Abyssinian Ground Hornbill	Great Hornbill	Rhinoceros Hornbill
Total genome length (bp)	1,159,178,761	1,132,614,623	1,104,977,869	1,065,782,792
Contig N50 (bp)	40,451,534	40,274,051	39,137,307	53,203
Contig N90 (bp)	13,418,008	14,657,043	13,223,265	12,582
Shortest contig (bp)	807	2,790	500	200
Longest contig (bp)	137,990,493	136,452,362	130,780,553	461,566
Number of contigs	1,674	113	20,477	62,257
Number of contigs larger than N50	9	9	9	5,769
Number of contigs larger than N90	31	29	30	21,237
Average transcript length (bp)	9,011	8,628	13,226	10,866
Average CDS length (bp)	214	223	190	175
GC content (%)	43.50	43.30	43.17	42.65
Total number of introns	175,151	172,647	160,400	130,435
Average number of introns per gene	6.83	6.91	7.70	6.29
Number of protein coding genes	17,646	17,392	16,422	18,387
Number of proteins identified by Augustus	33,746	33,774	26,676	25,137
Number of tRNAs	280	273	275	110
Number of rRNAs	50	13	22	4
Repeat content (%)	7.64	7.85	7.06	7.10

The completeness of the hornbill genomes and the quality of the predicted gene models were assessed using BUSCO with the avian lineage dataset composed of 8,338 genes. The GH genome had the greatest number of complete BUSCOs (96.28%), while the lowest was observed for the RH (89.1%) (Figure 3.1). The *Bucorvus* species, on average, achieved the highest completeness (95.73%), however, they did have a substantial number of missing genes (3.43%) compared to the *Buceros* species (1.66%).

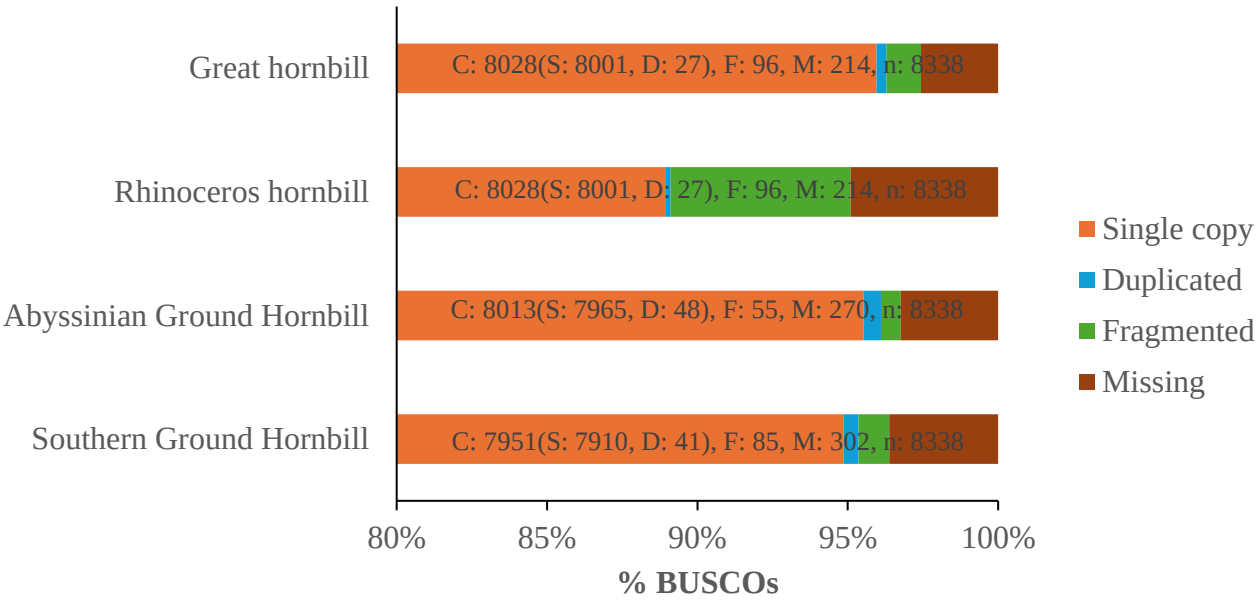


Figure 3.1: Assessment of genome assembly and annotation quality using BUSCO. BUSCO analysis was conducted with the Aves lineage dataset aves_odb10 which contained 8,338 genomes and 62 BUSCOs against which the hornbill genomes were assessed for completeness.

3.4.2 Structural analysis of the hornbill genomes

The hornbill genomes were predicted to harbour an average of 251 tRNAs and 32 rRNAs (Table 3.1). Similar numbers of tRNAs were predicted for all hornbills but the RH, which also had the lowest number of rRNAs. This could be due to the lower assembly quality of the RH, where this genome was assembled to 62,257 contigs with an N50 value of 53,203 bp. In contrast, the SGH assembly, with the highest N50 value also displayed the highest number of tRNAs (280) and rRNAs (50).

The 18S rRNA subunit forms part of the small ribosomal subunit which is essential for initiating protein synthesis (Doudna and Rath, 2002). It initiates the binding of ribosomes to mRNA molecules and ensures correct base pairing between tRNAs and mRNA during

translation (Locati *et al*., 2017). The rRNA subunits 5S, 5.8S and 28S form part of the large ribosomal subunit which contains the peptidyl transferase active site (Doudna and Rath, 2002). During protein synthesis, this subunit catalyses the formation of peptide bonds between amino acids (Doudna and Rath, 2002; Locati *et al*., 2017). Amongst the predicted rRNA subunits, the 5S and 28S rRNA subunits were predominant in the *Bucorvus* species (38.08% and 28.54%) compared to the *Buceros* species (34.09% and 26.14%), while the 18S rRNA subunit predominated in the *Buceros* (28.41%) compared to the *Bucorvus* species (22.54%). No 5.8S rRNAs were predicted in the RH genome.

tRNAs primarily carry the amino acids to the growing polypeptide chain in the ribosome allowing for the three base codons coding the protein sequence to be defined (Berg and Brandl, 2021). Within the four hornbill species, the tRNA carrying alanine was the most prevalent (11.24%) among all isotypes. The *Bucorvus* genomes had a greater proportion of tRNA isotypes carrying serine (9.77%), glycine (7.40%) and arginine (6.87%) compared to the *Buceros* species (5,55%, 6.91% and 5.18%, respectively) (Appendix Figure 1). In contrast, the *Buceros* species had a greater proportion of tRNA isotypes carrying lysine (8.82%) and leucine (7.36%) compared to the *Bucorvus* species (5.43% and 5.06%, respectively) (Appendix Figure 1). Furthermore, on average the *Bucorvus* genomes harboured approximately 3-fold more tRNA isotypes carrying glutamine and phenylalanine.

The repeat composition of the four hornbill genomes was assessed using RepeatMasker which indicated that on average 82.79 Mb (7.41%) of the hornbill genomes were comprised of repetitive elements (Table 3.2). The *Bucorvus* species had the highest repeat composition, with an average size of 88.74 Mb, which was 15.48% more than the *Buceros* species. The most prevalent repetitive elements were the class I retrotransposons (65.76%) followed by DNA transposons (13.44%). Class I retrotransposons function via a copy-and-paste mechanism using an RNA intermediate (Kapusta and Suh, 2017). Within the class I retrotransposons, all hornbill species had a greater proportion of long interspersed nuclear elements (LINEs - 57.99%), with the L2/CR1/Rex elements accounting for 99.77% of these elements. LINEs are autonomous retrotransposons that encode proteins required for retrotransposition via the target-primed reverse transcription mechanism (Kapusta and Suh, 2017). The CR1 elements, in particular, are the most predominant transposable elements identified in avian species (Kapusta and Suh, 2017). Short interspersed nuclear elements, SINEs, contributed 7.44% of the total repetitive elements on the hornbill genomes. SINEs are nonautonomous retrotransposons that lack the

required proteins required for retrotransposition and thus parasitize the enzymatic machinery of LINEs. These elements are derived from tRNAs, 5S rRNAs and other small RNAs and are therefore typically found in low copy numbers within avian species (Kapusta and Suh, 2017).

Table 3.2: Relative proportion of repetitive elements within the hornbill species

	Southern Ground Hornbill		**Abyssinian Ground Hornbill**		**Great Hornbill**		**Rhinoceros Hornbill**	
	Number of elements	% of sequence	Number of elements	% of sequence	Number of elements	% of sequence	Number of elements	% of sequence
Retroelements	186,729	65.38	187,447	66.13	176,503	65.46	176,573	66.07
SINEs:	52,684	7.31	52,915	7.25	48,386	7.49	48,166	7.72
Penelope	276	0.13	290	0.13	280	0.14	287	0.14
LINEs:	131,842	57.82	132,323	58.50	125,874	57.56	126,205	58.07
L2/CR1/Rex	131,309	57.69	131,765	58.38	125,353	57.42	125,684	57.93
L1/CIN4	257	0.00	268	0.00	241	0.00	234	0.00
LTR	2,203	0.26	2,209	0.25	2,243	0.28	2,202	0.28
Gypsy/DIRS1	313	0.13	297	0.13	299	0.14	303	0.00
DNA transposons	103,581	12.95	104,225	13.00	98,684	13.73	99,209	14.07
hobo-Activator	25,382	2.69	25,600	2.75	24,683	2.91	24,734	3.03
Tc1-IS630-Pogo	4,553	0.51	4,538	0.50	4,404	0.55	4,415	0.55
Tourist/Harbinger	26,386	4.10	26,637	4.13	23,758	4.16	23,894	4.28
Rolling-circles	588	0.13	599	0.13	538	0.14	600	0.14
Unclassified:	47,596	3.97	47,927	4.00	47,107	4.44	47,210	4.55
Total interspersed repeats		**82.31**		**83.13**		**83.63**		**84.69**
Small RNA	70,501	5.26	70,637	5.25	67,520	5.41	65,681	5.38
Satellites	0	0.00	0	0.00	2	0.00	3	0.00
Simple repeats	205,915	10.26	205,055	9.50	179,150	8.88	160,732	8.00
Low complexity	40,633	2.18	40,597	2.13	35,948	2.08	33,039	1.93
Total	**655,543**	**100.00**	**656,487**	**100.00**	**605,452**	**100.00**	**583,047**	**100.00**

3.4.3 Pan-genome analysis of the hornbills

The pan-genome of the four hornbill species was comprised of a total of 119,333 proteins assigned to 40,294 orthogroups. The core genome shared amongst all hornbill species comprised 11,364 orthogroups (28.20% of the pan-genome), of which 72.13% were SCOs (Figure 3.2). The accessory genome comprised 28,930 orthogroups, of which 48.88% were species-specific proteins. The highest proportion of species-specific proteins (29.00% - 4,100 proteins) was observed in the GH, while the AGH had the lowest proportion (21.98% - 3,108 proteins). The *Bucorvus* species, however, did share the largest proportion of the accessory genome (24.33% - 7,038 orthogroups), while 3,937 orthogroups were shared by the *Buceros* species only. The *Bucorvus* species individually shared less than 1.00% of the accessory genome with either of the *Buceros* species (Figure 3.2).

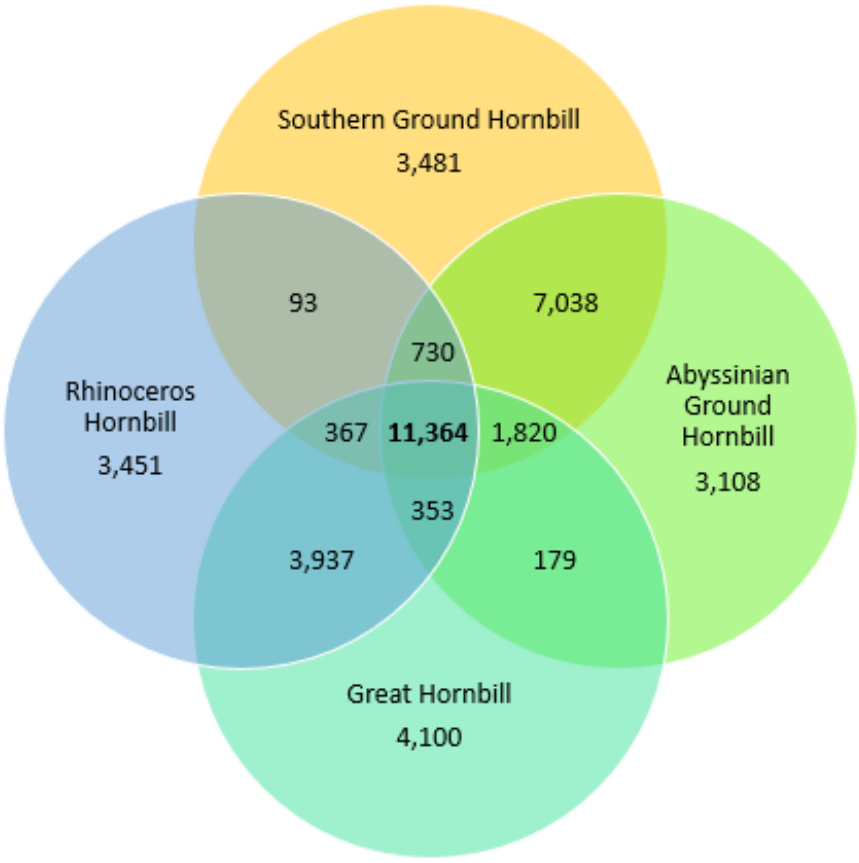


Figure 3.2: Pan-genome analysis of the four hornbill species. Venn diagram shows the distribution of the core genome shared by all species and the accessory genome which includes species-specific orthogroups and those shared by one or more species but not all. The core and accessory genomes contain the total number of orthogroups which includes both SCOs and paralogues proteins.

The core and unique orthogroups were assigned to their COG supra-functional categories ("information storage and processing", "cellular processes and signalling", "metabolism", and "poorly characterized functions") and twenty-seven COG categories. Approximately 96% of the core orthogroups and 64% of the species-specific orthogroups (15,389 proteins) had a functional COG category assignment.

Of the core orthogroups, 33.12% (5,700 orthogroups) belonged to the "poorly characterised function" supra-functional category (Figure 3.3). Core orthogroups with a defined function were predominantly associated with the supra-functional category "cellular processes and signalling" (36.36%). Within this COG supra-functional category, most proteins were involved in signal transduction (18.62%) and post-translational modification, protein turnover and chaperone functions (6.43%) (Appendix Table 5). Core orthogroups involved in transcription, within the COG supra-functional category of "information storage and processing", were also prevalent among the hornbills (7.07%).

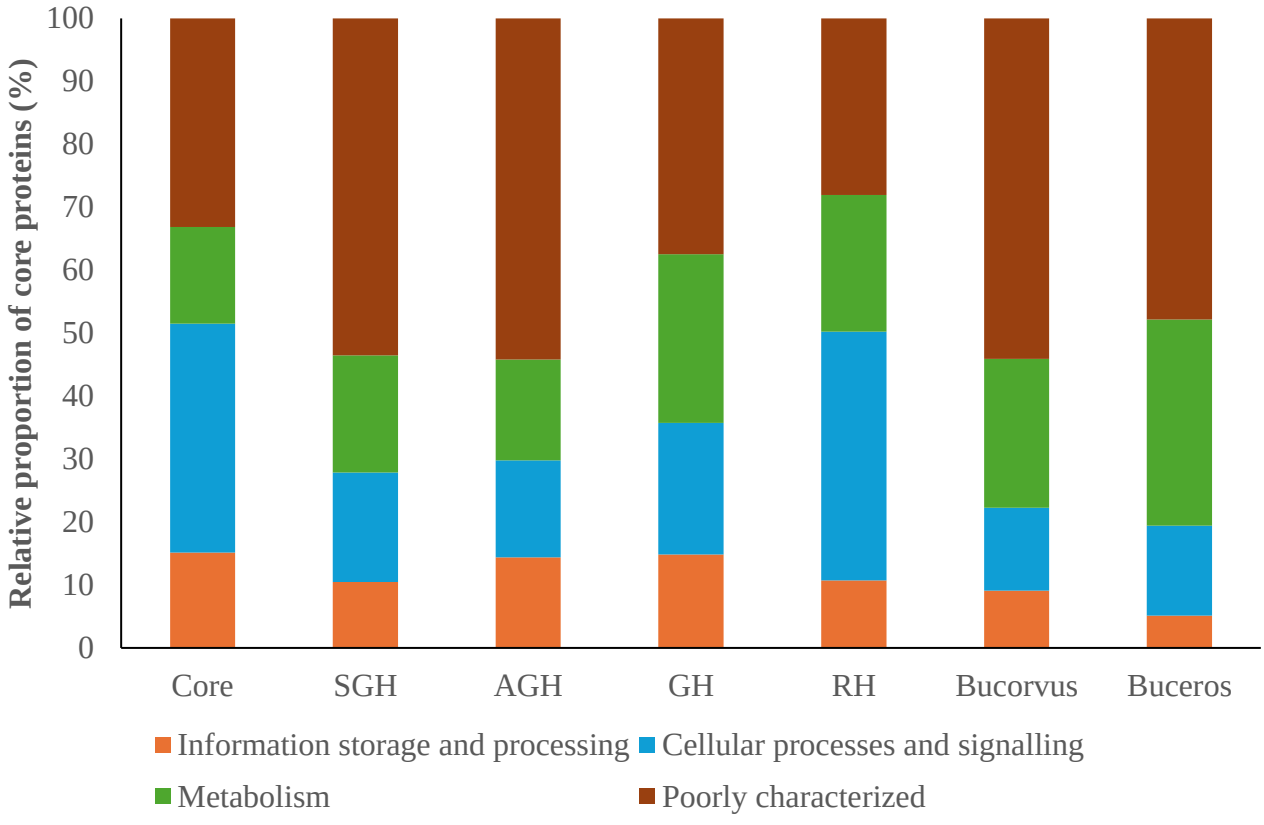


Figure 3.3: Relative proportion of the core and accessory genome of the four hornbill species in each COG super-functional group. The core genome represents the shared orthogroups between all four hornbill species, followed by the species-specific orthogroups in the accessory genome and the shared orthogroups between the two hornbill genera.

On average, 39.89% of the species-specific proteins of the four hornbill species had a "poorly characterised function", with a greater proportion of these encoded on the genomes of members of the genus *Bucorvus* (53.85%) than those of the *Buceros* species (32.74%) (Figure 3.3). The species-specific proteins with a defined function in the *Bucorvus* species were mainly associated with the COG supra-function category "metabolism" (17.32%), followed by "cellular processes and signalling" (16.42%). By contrast, the genomes of the *Buceros* species encoded species-specific proteins associated with "cellular processes and signalling" (30.91%), followed by "metabolism" (24.29%). Within the COG supra-function category "cellular processes and signalling", the *Buceros* species genomes incorporated 26-fold and 11-fold more species-specific proteins associated with the COG categories cell motility and cell wall/membrane/envelope biogenesis, respectively, than the *Bucorvus* species (Appendix Table 6). Furthermore, the *Buceros* species also displayed 5-fold more species-specific proteins associated with the COG category coenzyme metabolism within the COG supra-function category metabolism compared to the *Bucorvus* species.

Apart from proteins associated with a "poorly characterised function" (50.96%), the *Buceros* species shared more proteins associated with the COG supra-functional category "metabolism" (32.74%) and "cellular processes and signalling" (14.25%) between them than the two *Bucorvus* species (23.66% and 13.17%, respectively) (Figure 3.3). At the COG category level, the two *Bucorvus* species shared 4-fold more proteins associated with energy production and conversion and 3.5-fold more proteins associated with cell cycle control and mitosis and cell wall/membrane/envelope biogenesis compared to the *Buceros* species (Appendix Table 6). Additionally, the two *Bucorvus* species shared more proteins associated with the COG supra-functional category "information storage and signalling" (9.11%) compared to the *Buceros* species (5.15%). Within this COG supra-functional category, the *Bucorvus* species shared 7.5-fold and 4-fold more proteins associated with DNA replication and repair and nuclear structure compared to the *Bucorvus* species.

3.4.4 Functional analysis of the accessory genome using KEGG pathways

A total of 7,038 orthogroups are shared between the *Bucorvus* species only, while 3,937 are shared by the *Buceros* species but absent from the *Bucorvus* species. These shared orthogroups excluded the core orthogroups shared by all species and the species-specific orthogroups. KO terms were assigned to 23.39% (3,911 proteins) of the orthogroups shared by the *Bucorvus* and those by the *Buceros* species. These KO terms were used to reconstruct the KEGG pathways

in which these proteins function. It should be noted that many KO terms are associated with more than one pathway. Thus, several of the proteins in *Bucorvus* and *Buceros* shared orthogroups were also associated with more than one function. For example, proteins with KO terms associated with the class "human diseases" were also linked to other major KEGG classes such as "metabolism" and "organism systems". The KEGG pathways that exhibited a five-fold difference in the number of proteins associated with the shared orthogroups of either the *Bucorvus* species or the *Buceros* species were further analysed to identify functional differences between the two genera.

3.4.4.1 Lipid and energy metabolism

A total of 378 proteins were associated with the KEGG class "metabolism". Of these, 250 proteins were encoded on the genomes of the AGH and SGH, while only 128 proteins were encoded on the genomes of the GH and RH. Analysis of the individual KEGG pathways within "metabolism" showed that the *Bucorvus* species shared approximately 8-fold more proteins involved in lipid metabolism than the *Buceros* species (Appendix Table 7). Lipid metabolism is a complex process that involves several biochemical processes functioning in the synthesis, breakdown, storage and utilisation of lipids in organisms (Gyamfi *et al*., 2019). These molecules are essential for energy production and the synthesis of structural and functional lipids involved in the construction of cellular components and signalling (Chandel, 2021; Gyamfi *et al*., 2019). Proteins involved in all thirteen of the pathways involved in lipid metabolism were observed in the *Bucorvus* species, while the *Buceros* species lacked proteins involved in ten of the pathways (Appendix Table 8).

Fatty acid biosynthesis allows for the conversion of excess dietary carbohydrates and amino acids into fatty acids. These fatty acids can be used to synthesise triacylglycerols (TAGs) in the liver and adipose tissues of avian species via the glycerolipid metabolism pathway (Gyamfi *et al*., 2019; Hermier, 1997). These deposits are then used for energy and allow the body to survive during periods of fasting and starvation (Araújo *et al*., 2019; Gyamfi *et al*., 2019). The *Bucorvus* genomes encode five proteins associated with the glycerolipid metabolism pathway, while those of the *Buceros* species only encode one protein. (Figure 3.4; Appendix Table 8).

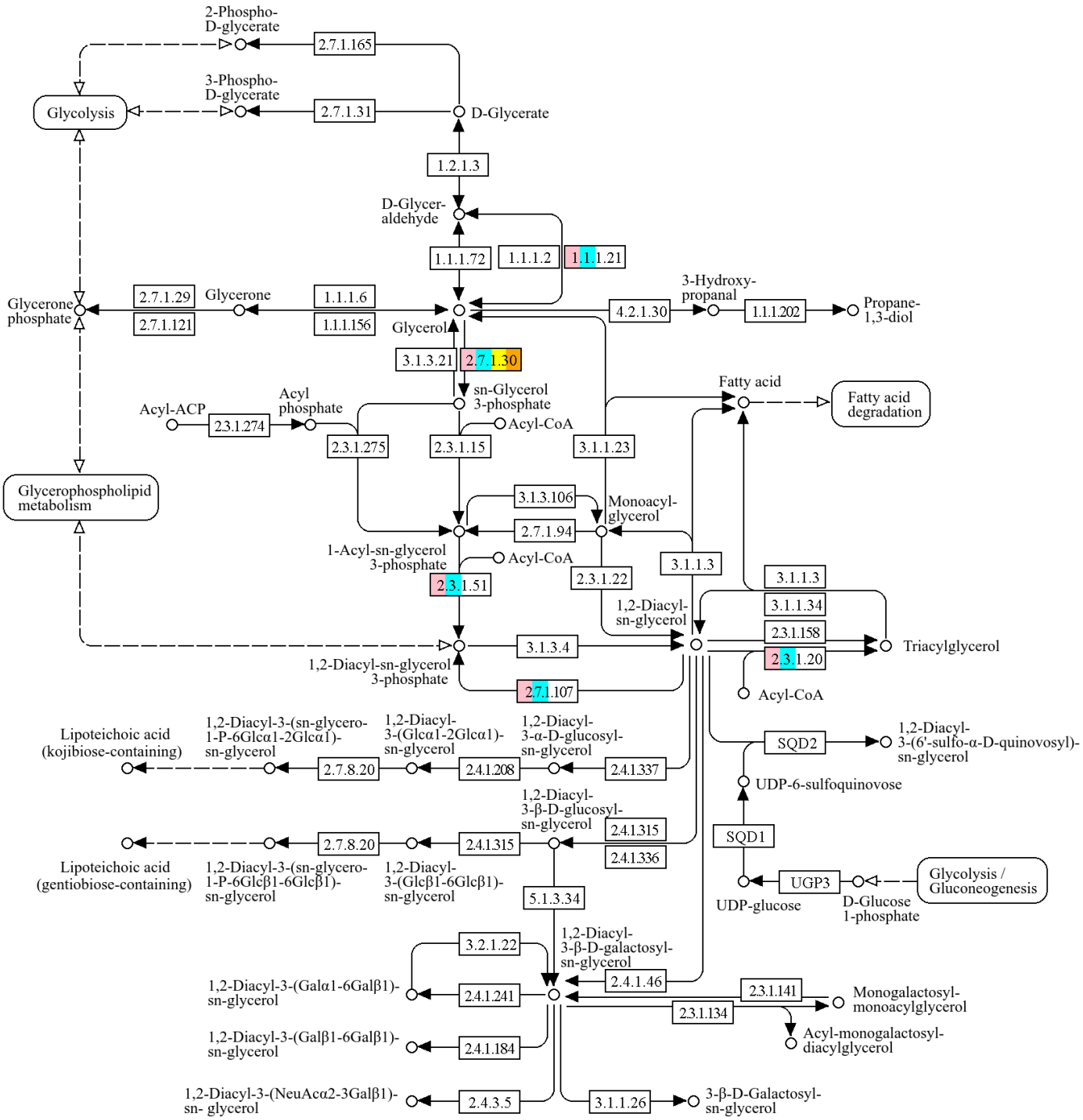


Figure 3.4: Proteins involved in the glycerolipid metabolism pathway. Proteins identified in the shared orthogroups of the *Bucorvus* species are indicated in pink (SGH) and cyan (AGH) while those identified in the *Buceros* species are indicated in yellow (GH) and orange (RH). The identified enzymes include aldehyde reductase (EC: 1.1.1.21), glycerol kinase (EC: 2.7.1.30), diacylglycerol kinase (ATP – EC: 2.7.1.107), diacylglycerol, O-acyltransferase 1 (EC: 2.3.1.20) and lysophospholipid acyltransferase 1/2 (EC: 2.3.1.51).

The *Bucorvus* genomes also code for six proteins associated with the glycerophospholipid metabolism pathway which were absent in the *Buceros* species. This pathway is essential for the synthesis of membrane phospholipids, including glycerophospholipids and sphingolipids (Dai *et al*., 2021). These phospholipids play a role in forming a barrier that separates cells from their environment and different cellular compartments (Dai *et al*., 2021).

Apart from lipid metabolism, the genomes of the two *Bucorvus* species also encode 5-fold more proteins linked to energy metabolism compared to the *Buceros* species (Appendix Table 7). Within energy metabolism, ten shared proteins associated with oxidative phosphorylation were encoded on the genomes of the *Bucorvus* species, compared to the *Buceros* species which only encoded one (Figure 3.5). Energy metabolism is the process of generating energy in the form of adenosine 5′-triphosphate (ATP) from nutrients such as carbohydrates, proteins and fats (Yang *et al*., 2020). The process of oxidative phosphorylation is the primary pathway for generating ATP from these nutrients via the electron transport chain (ETC) in the mitochondrion (Nolfi-Donegan *et al*., 2020; Wilson, 2017).

The lipid and energy metabolism proteins unique to either the genus *Bucorvus* or *Buceros* could potentially reflect differences in various aspects of their metabolism and dietary habits in response to their respective environments (Isaksson *et al*., 2017; Qu *et al*., 2013). For instance, the higher number of proteins associated with oxidative phosphorylation in the *Bucorvus* species could indicate that they have greater energy requirements for survival and functioning. These energy requirements could be related to processes such as thermogenesis (process of heat production in response to changing environmental temperatures) which might be more demanding than those of the *Buceros* species.

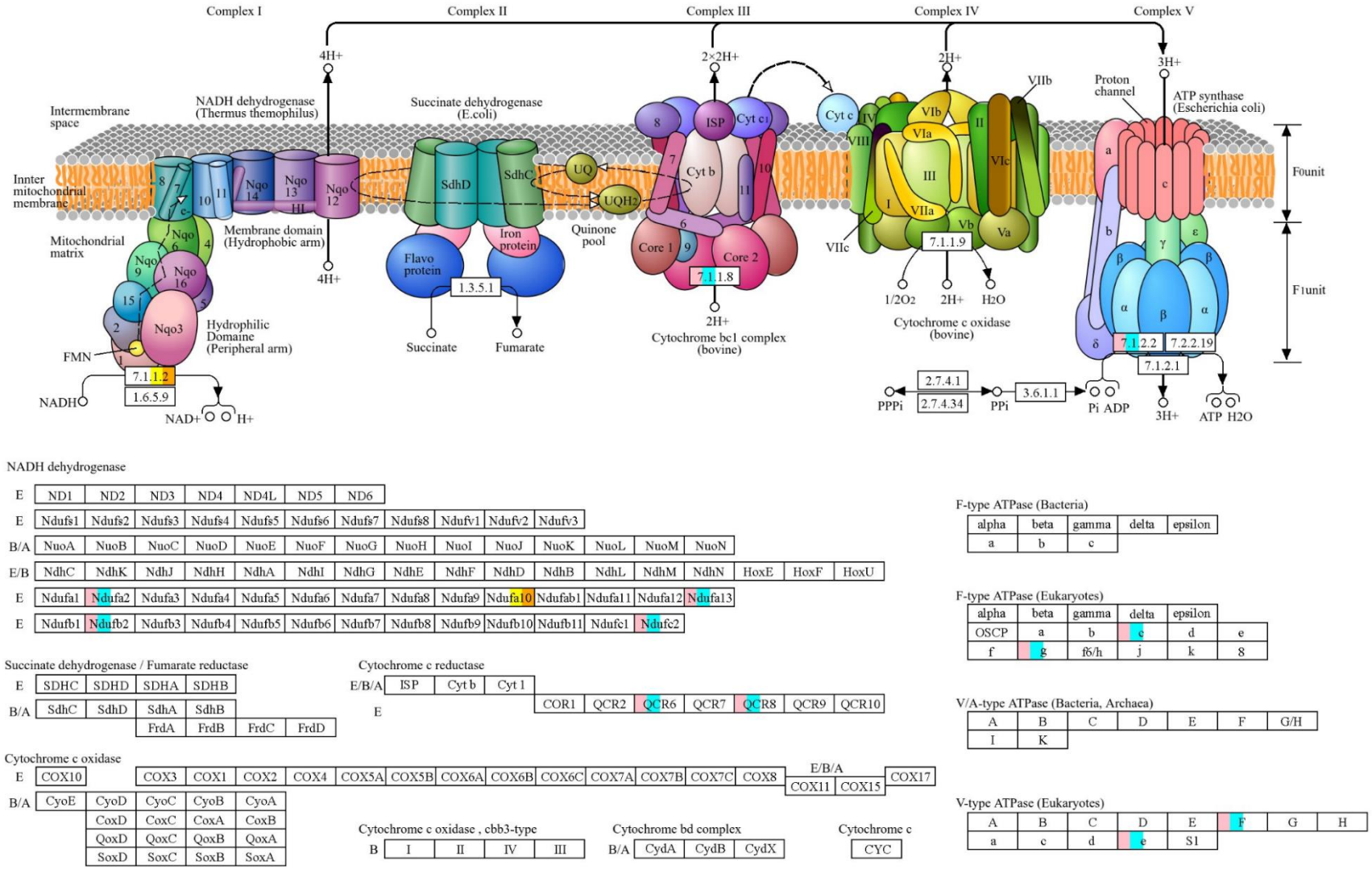


Figure 3.5: Proteins involved in the oxidative phosphorylation pathway. Proteins identified in the orthogroups shared by the *Bucorvus* species are indicated in pink (SGH) and cyan (AGH) while those identified in the *Buceros* species are indicated in yellow (GH) and orange (RH). The identified enzymes include ubiquinol-cytochrome c reductase (QCR) subunit 6 (QCR6), QCR subunit 8 (QCR8), F-type H+-transporting ATPase (ATPeF) subunit c (ATPeF0C), ATPeF subunit g (ATPeFG), V-type H+-transporting ATPase (ATPeV) subunit F (ATPeV1F), ATPeV subunit e (ATPeV0E), NADH dehydrogenase (ubiquinone) 1 alpha (NADUFA) subcomplex subunit (NDUFA2), NADUFA subcomplex subunit (NDUFA10), NADUF beta subcomplex subunit 2 (NDUFB2), NADUF subunit C2 (NDUFC2) and NADUFA subcomplex subunit (NDUFA13).

3.4.4.2 Organism systems

A total of 533 proteins were associated with the KEGG class "organism systems", of which 322 proteins were encoded on the genomes of the SGH and AGH. In contrast, only 201 proteins were encoded on the genomes of the GH and RH. The organism systems represent a complex interplay of several organ systems such as the immune system, endocrine system, digestive system and excretory system, as well as systems for developmental and regeneration, and environmental adaptation required by avian species to meet the demands of their diverse lifestyles (Scanes, 2020).

The excretory system

The *Bucorvus* species exhibited 7-fold more shared proteins linked to the excretory system compared to the *Buceros* species (Appendix Table 7). This organ system incorporated five KEGG pathways of which four pathways were specific to the *Bucorvus* species (Appendix Table 9). These include the aldosterone-regulated sodium reabsorption, proximal tubule bicarbonate reclamation and collection of duct acid secretion pathways, which play an important role in the homeostasis of sodium chloride via aldosterone secretion, bicarbonate and acid-base byproducts in avian species (DuBose, 1990; Hughes, 2003; Wagner *et al*., 2009). Furthermore, the *Bucorvus* species genomes also coded for shared proteins functioning in the vasopressin-regulated water reabsorption pathway which were absent in the *Buceros* species (Appendix Table 9). Vasopressin-regulated water reabsorption maintains water balance using vasopressin, an antidiuretic hormone (Goldstein, 2006). This hormone is expressed at high levels in avian species during dehydration to increase the reabsorption of water from urine back into the bloodstream (Goldstein, 2006). Furthermore, variations in diet, habitat aridity and lifestyle have been identified to affect osmoregulation physiology in several avian species (Fleming and Nicolson, 2003; Guglielmo *et al*., 2017; Sabat *et al*., 2009; Song and Beissinger, 2020). Therefore, the more complex excretory proteome identified in the *Bucorvus* species could have evolved to regulate these factors more effectively. This adaption could have been in response to the variation in diet (faunivorous vs frugivorous), lifestyle (terrestrial vs arboreal) and habitat between the *Bucorvus* and *Buceros* species, respectively.

The immune system

Within the immune system, the *Bucorvus* species exhibited 6-fold more shared proteins linked to the NOD-like receptor (NLR) signalling pathway compared to the *Buceros* species. The NLR pathway plays a crucial role in the innate immune system, the first line of defence that recognises microbial components of invading pathogens. This pathway initiates an inflammatory and antimicrobial response within avian species (Ma *et al*., 2021). Unique to the *Bucorvus* were five proteins associated with the biosynthesis of interleukins, interferons and cytokines (Appendix Table 10).

The genomes of the *Bucorvus* species also encoded ~ 5.5- and 5-fold more shared proteins linked to Th17 cell differentiation and IL-17 signalling pathways, respectively, compared to the *Buceros* species. The Th17 cell differentiation pathway promotes the differentiation of T helper (Th)-17 cells, a subset of $CD4^+$ T cells that assist in the production of IL-17 cytokines (Harrington *et al*., 2005; Park *et al*., 2005). The IL-21 receptor and IL-23 subunit alpha proteins, unique to the *Bucorvus* species, play an important role in promoting this differentiation of naïve $CD4^+$ T cells into Th17 cells (Earnshaw and Warren, 2023; Korn *et al*., 2007; Tesmer *et al*., 2008). The binding of IL-17 cytokines to receptors on target cells activates the IL-17 signalling pathway, which in turn results in an inflammatory response via the expression of cytokine and chemokines (Mills, 2023). The *Bucorvus* genomes code for five proteins associated with this pathway, while those of the *Buceros* species only code for one protein (Appendix Table 10).

Th17 cells and IL-17 chemokines function primarily in the adaptive immune response, which is a more specific immune response compared to the innate immune response (Isailovic *et al*., 2015). Similarly, the B cell receptor signalling pathway, with 5-fold more shared proteins present in the *Bucorvus* species genomes compared to the *Buceros* species, also plays a role in the adaptive immune response (Appendix Table 10) (Xie *et al*., 2023). This pathway results in the activation, differentiation and functioning of B cells which produce antibodies (Roberts and Taraska, 2023). Within the *Bucorvus* species, five proteins were predicted to function in this pathway that were absent in the *Buceros* species. Studies have shown that investments in immune response in avian species are modulated based on pathogen pressure (Hegemann *et al*., 2019, 2012; Horrocks *et al*., 2015). Thus, the variation in the immune system proteome in the *Bucorvus* and *Buceros* species could reflect the differences in the kind of pathogens these genera encounter in their respective environments, as well as their adaptive responses to these exposures.

The endocrine system

In contrast to the excretory system and the immune system, in terms of the endocrine system, the genomes of the *Buceros* species code for six proteins involved in the insulin secretion pathway, while only one is encoded on those of the *Bucorvus* species (Figure 3.6; Appendix Table 11). This pathway helps regulate blood sugar levels by producing insulin, primarily from pancreatic beta cells, which create insulin-filled granules that release insulin when triggered by glucose or other stimuli (Hoang Do and Thorn, 2015). Unlike mammals, avian species rely less on insulin and instead maintain blood glucose levels through insulin resistance and elevated plasma glucose levels (Braun and Sweazea, 2008). In these species, the role of insulin is more pronounced in the fed state, where it helps manage postprandial glucose levels, while glucagon plays a more critical role during fasting (Ji *et al*., 2020; Tokushima *et al*., 2005). While the *Bucorvus* species did exhibit a lower number of proteins involved in the insulin secretion pathway they did have 2-fold more shared proteins functioning in the insulin and glucagon signalling pathways compared to the *Buceros* species.

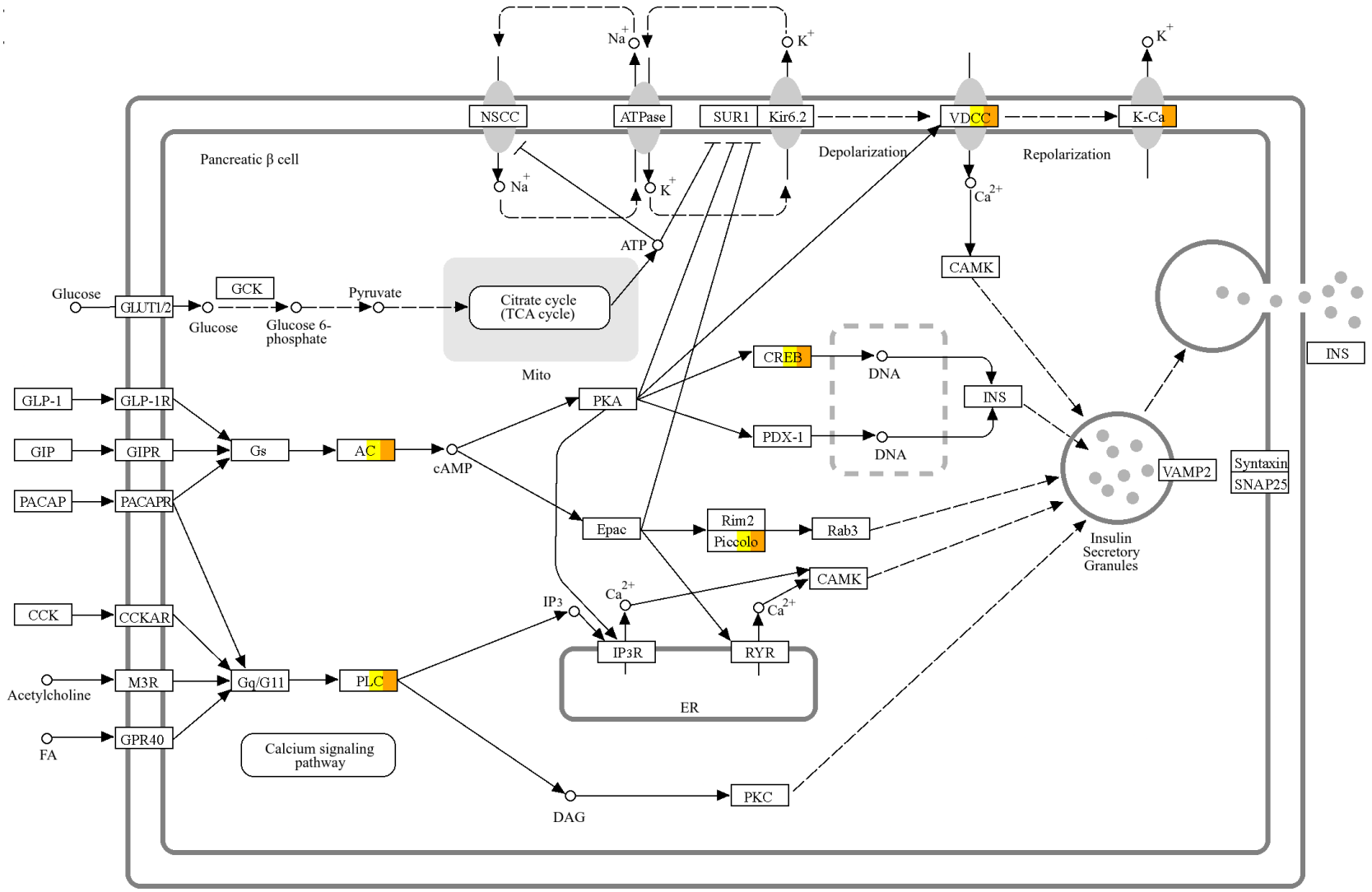


Figure 3.6: Proteins involved in the insulin secretion pathway. Proteins identified in the orthogroups shared by the *Bucorvus* species are indicated in pink (SGH) and cyan (AGH) while those identified in the *Buceros* species are indicated in yellow (GH) and orange (RH). The identified enzymes included cyclic AMP-dependent transcription factor ATF-4 (ATF4), voltage-dependent calcium channel L type alpha-1D (CACNA1D), potassium large conductance calcium-activated channel subfamily M alpha member 1 (KCNMA1), phosphatidylinositol phospholipase C, beta (PLCB), adenylate cyclase 2 (ADCY2 – EC: 4.6.1.1).

The digestive system

The genomes of the *Bucorvus* species code for three proteins functioning in protein digestion and absorption pathway in the digestive system. Among the *Buceros* species, only one protein functioning in this pathway was present (Appendix Table 11). Dietary protein is broken down and digested first in the lumen by gastric and pancreatic proteases followed by digestion on the external surface of membranes by peptidases (Bhutia and Ganapathy, 2018). This results in the production of amino acids and peptides which are then transported into the intestinal epithelium cells (Bhutia and Ganapathy, 2018). In avian species, protein digestion and absorption are essential for maintaining structural integrity, enzymatic activity, immune responses and promoting growth (Bhutia and Ganapathy, 2018; Selle and Liu, 2019).

Osteoclast differentiation

The osteoclast differentiation pathway is characterised under the KEGG category of development and regeneration. Osteoclasts are multinucleated cells that originate from the hematopoietic system, which produces blood cells and is responsible for reabsorbing old bone matrix (Kim and Kim, 2016). The pathway is responsible for the differentiation of osteoclasts from monocyte or macrophage cells, which is essential for maintaining bone homeostasis and haematopoiesis (formation of cellular components associated with blood) (Hu *et al*., 2022; Kim and Kim, 2016). A total of seven shared proteins functioning in the osteoclast differentiation pathway were identified in the *Bucorvus* species genomes that were absent in the *Buceros* species (Appendix Table 12). This indicates that these two genera could have evolved differences in the bone remodelling processes, potentially leading to differences in their skeletal structures, bone health, or adaptation to their respective environments based on their lifestyles (terrestrial vs arboreal lifestyle).

Thermogenesis

Being endotherms, avian species maintain high body temperatures, necessitating that a large proportion of their energy budget is utilised to maintain these temperatures (Hohtola, 2012). Birds utilise thermogenesis to generate the necessary heat as a by-product of metabolic activity in the body when temperatures decrease below thermoneutrality via shivering or non-shivering mechanisms (Yau and Yen, 2020). Furthermore, thermogenesis enables birds to adapt to temperature fluctuations in cold environments (Pani and Bal, 2022). The *Bucorvus* genomes

code for thirteen proteins involved in thermogenesis compared to the *Buceros* species with only three proteins for this function (Appendix Table 12). This could reflect the changes in environmental conditions and adaptive mechanisms with varying thermogenic requirements between the genus *Bucorvus* and *Buceros* (savannah biomes vs tropical forests).

3.4.4.3 Genetic information processing

A total of 148 proteins were associated with the KEGG class "genetic information processing" of which 102 proteins were encoded on the genomes of the *Bucorvus* species compared to the 46 on the *Buceros* species. Within this class, the genomes of the *Bucorvus* species encode 5.5-fold more proteins linked to the spliceosome compared to the *Buceros* species. The spliceosome is a large RNA-protein complex, comprised of five small nuclear ribonucleoproteins (snRNP - U1, U2, U4, U5 and U6) and several protein cofactors, that play a role in excising introns present in the mRNA, essential for effective gene expression (Lamond, 1993; Matera and Wang, 2014). A total of twelve proteins were exclusively encoded on the genomes of the *Bucorvus* species, which mainly played a role in forming complexes required for the assembly of the spliceosome (Figure 3.7) (Ayadi *et al*., 1998; Kondo *et al*., 2015; Nguyen *et al*., 2015). In contrast, the two proteins encoded on the *Buceros* genomes were associated with splicing factors (Albers *et al*., 2003; Corioni *et al*., 2011; Figueroa and Hayman, 2004; Goldstrohm *et al*., 2001).

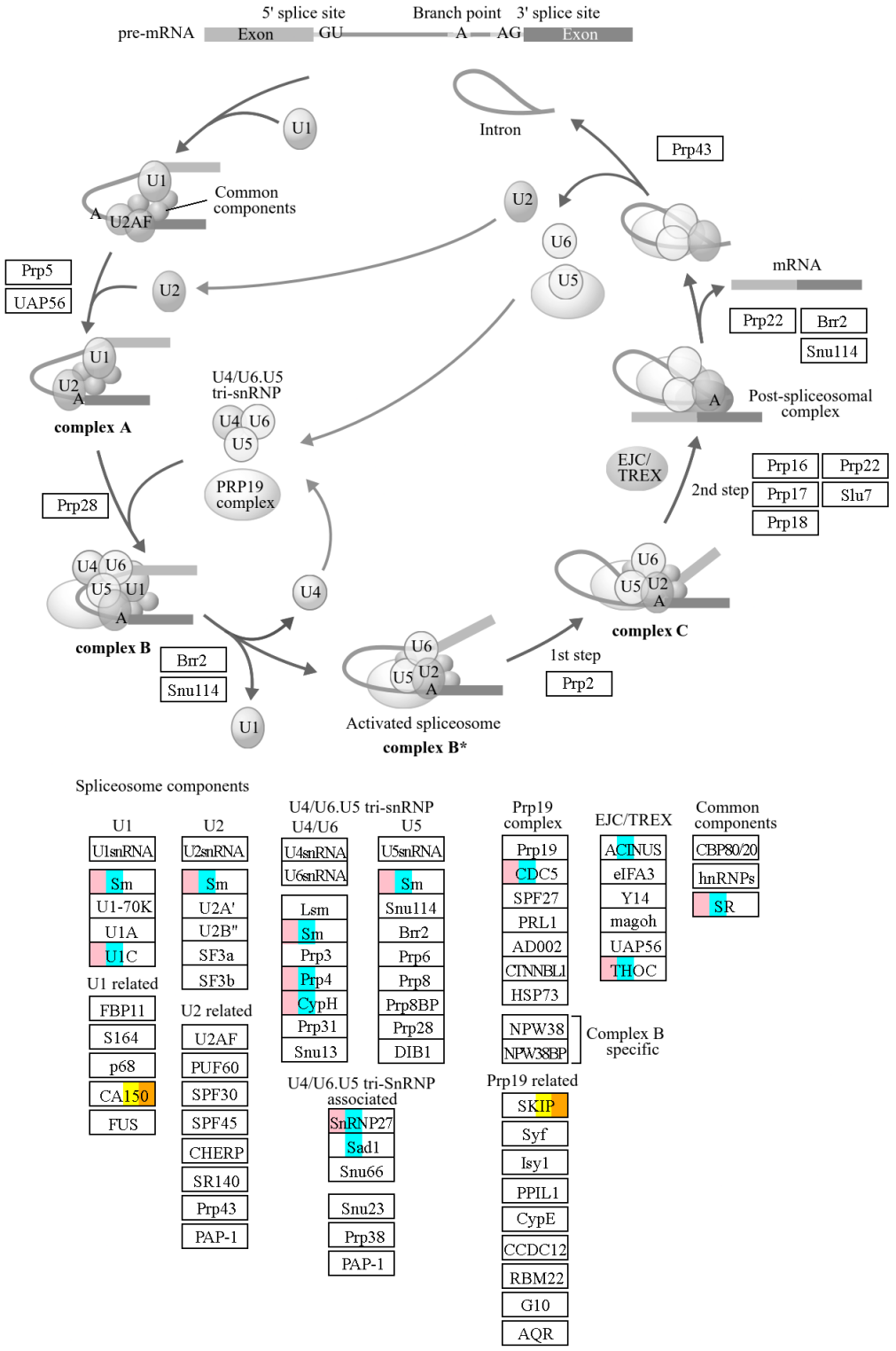


Figure 3.7: Proteins involved in the spliceosome. Proteins identified in the orthogroups shared by the *Bucorvus* species are indicated in pink (SGH) and cyan (AGH) while those identified in the *Buceros* species are indicated in yellow (GH) and orange (RH). The identified enzymes included SNW domain-containing protein 1 (SNW1), peptidyl-prolyl isomerase H (cyclophilin H) (PPIH), U1 small nuclear ribonucleoprotein C (SNRPC), small nuclear ribonucleoprotein D2 (SNRPD2), small nuclear ribonucleoprotein G (SNRPG), U4/U6 small nuclear ribonucleoprotein PRP4 (PRPF4), transcription elongation regulator 1 (TCERG1), U4/U6.U5 tri-snRNP-associated protein 3 (SNRNP27), U4/U6.U5 tri-snRNP-associated protein 2 (USP39), pre-mRNA-splicing factor CDC5/CEF1 (CDC5L), apoptotic chromatin condensation inducer in the (ACIN1), THO complex subunit 4 (THOC4), serine/arginine-rich splicing factor 1 (SRSF1) and serine/arginine-rich splicing factor 7 (SRSF7).

3.4.4.4 Environmental information processing

Cells within organisms communicate and respond to external stimuli through the intricate interplay of signalling molecules, receptors and intracellular pathways. This complex system allows cells to perceive and correctly respond to their microenvironment, enabling them to coordinate the necessary biological responses (Cooper, 2000; Valls and Esposito, 2022). A total of 187 proteins linked to the category signal transduction were encoded on the genomes of the *Bucorvus* species compared to the 82 encoded on the *Buceros* species genomes.

In particular, the *Bucorvus* species shared eight proteins associated with the Janus kinase-signal transducer and activator of transcription (JAK-STAT) signalling pathway that were absent in the *Buceros* species (Appendix Table 13**).** A total of ten proteins involved in the Ras signalling pathway are encoded on the genomes of both *Bucorvus* species, while only two were coded for on the *Buceros* genomes. The JAK-STAT and Ras signalling pathways play roles in transmitting signals to the nucleus for the regulation of cell proliferation, differentiation, survival, migration and apoptosis (Harrison, 2012; Molina and Adjei, 2006; Slack, 2017).

A total of five shared proteins functioning in the adenosine monophosphate-activated protein kinase (AMPK) signalling pathway were identified in the genomes of the *Bucorvus* species, while only one protein was present in the genomes of the *Buceros* species (Appendix Table 13). AMPK is a serine-threonine protein kinase enzyme complex that monitors cellular energy via AMP, adenosine diphosphate (ADP) and ATP levels (Aslam and Ladilov, 2022). Low levels of ATP and high levels of AMP trigger the AMPK signalling pathway to restore the energy balance by restricting anabolic processes consuming ATP and promoting catabolic processes that generate ATP (Aslam and Ladilov, 2022; Hardie, 2014). In avian species, the AMPK signalling pathway has been implicated in energy homeostasis and nutrient regulation (Hu *et al*., 2019; Wang *et al*., 2020). In particular, high-energy diets inhibit the AMPK signalling pathway and suppress appetite, while low-energy diets activate the AMPK signalling pathway and increase appetite (Hu *et al*., 2019).

3.4.4.5 Cellular processes

Oocyte meiosis is the process through which female germ cells are differentiated into mature oocytes or eggs. Oocytes undergo two consecutive divisions via meiosis I and II to produce mature oocytes (Alberts *et al*., 2002a). In avian species, oocytes are arrested in prophase I of meiosis and enter a growth period during which they accumulate the components required for embryonic development (Schneider, 2009). Once sexual maturity is reached, via the stimulation of hormones, the cells resume cell division of meiosis I (Alberts *et al*., 2002a). Within the class "cellular processes", a total of five shared proteins were identified in the *Bucorvus* species functioning in the oocyte meiosis pathway (Appendix Table 14). In comparison, the genomes of the *Buceros* species coded for only one protein functioning in this pathway.

3.4.5 Gene duplication events

Gene duplication serves as an important mechanism in evolution which allows for the emergence of new genes and functions (Magadum *et al*., 2013). Although every gene has an equal probability of duplication, not all duplicates are preserved equally (McGrath *et al*., 2014). Most newly duplicated genes are silenced shortly after their formation (Lynch and Conery, 2000). However, those that are retained can either maintain their original function or evolve to acquire new ones. This dynamic process of gene duplication significantly contributes to species adaptation to diverse environments (Magadum *et al*., 2013). In certain instances, it also facilitates species diversification and increases biological complexity between species (Van de Peer *et al*., 2009).

The core orthogroups were analysed to identify paralogous proteins that may have been derived through gene duplication events in the *Buceros* and *Bucorvus* species. More than a quarter (3,167 orthogroups – 27.87%) of the orthogroups were identified to incorporate paralogous proteins within either the *Bucorvus* or *Buceros* species or in both genera. A total of 42 orthogroups, comprising 3,041 proteins, were further identified to have over 5-fold more protein copies in one genus compared to the other. Amongst the paralogous proteins, the genomes of the *Bucorvus* species coded for 2,727 paralogous proteins, which is 8.68 times more than that coded in the genomes of the *Buceros* species (314 paralogous proteins).

The functional implications of paralogous proteins derived through gene duplication in the *Bucorvus* and *Buceros* species were determined based on their COG category assignment and KO terms. The majority (95.10%) of proteins within the orthogroups with 5-fold more protein copies in either the genus *Bucorvus* or *Buceros* were assigned a functional COG category. Overrepresented paralogues in both the genus *Bucorvus* and *Buceros* were predominated by those belonging to the COG supra-functional category "poorly characterised function" with 63.99% and 49.80%, respectively. The *Buceros* species have a greater proportion of overrepresented proteins involved in the COG supra-functional categories "metabolism" (19.66%), "information storage and processing" (21.13%) and "cellular processes and signalling" (9.40%) compared to the *Bucorvus* species (8.51%, 20.18% and 7.32%, respectively).

The paralogous proteins overrepresented in the *Buceros* species primarily belonged to the COG category amino acid metabolism and transport (18.51%), whereas it only contributed 7.82% of the *Bucorvus* overrepresented paralogue dataset. On average proteins involved in DNA replication and repair (13.36%), transcription (6.15%) and post-translational modification, protein turnover and chaperone functions (6.14%) were overrepresented in both the *Bucorvus* and *Buceros* species.

Approximately 15% of proteins within the orthogroups with 5-fold more protein copies in either the genus *Bucorvus* or *Buceros* had KO term assignments, which were used to construct the KEGG pathways. This revealed that 50.13% and 53.23% of the overrepresented proteins in the *Bucorvus* and *Buceros* species were linked to the class "human diseases", respectively. The genomes of the *Bucorvus* species also encode more paralogous proteins linked to the classes "organism systems" (20.32%) and "environmental information processing" (13.19%) compared to the *Buceros* species (17.91% and 12.44%).

Analysis of the reconstructed KEGG pathways indicated that paralogous proteins linked to the excretory systems were found at a 4-fold greater abundance, while those associated with the KEGG categories 'nucleotide metabolism', 'translation', 'protein folding and degradation', and 'signalling molecules and interaction', were found at a 3-fold greater abundance in the *Bucorvus* species compared to the *Buceros* species (Appendix Table 7).

The *Bucorvus* species exhibited 5-fold more paralogous proteins linked to pathways under the category 'cellular community' compared to the *Buceros* species. This category includes the focal adhesion and tight junction pathways as well as signalling pathways regulating the pluripotency of stem cells. The *Bucorvus* species genomes coded for three paralogous proteins associated with the focal adhesion pathway, which were absent in the *Buceros* species (Appendix Table 15). The focal adhesion pathway is a signalling pathway that functions in the production, maintenance and regulation of focal adhesions, which are specialised structures that link the extracellular matrix (ECM) to the cell (Wu, 2007). This allows for the transmission of signals bidirectionally between the ECM and the cell (Miller *et al*., 2020).

In the four hornbill species, one orthogroup comprising 1,923 paralogous copies was substantively over-represented in the *Bucorvus* (average 831.5 copies/genome) compared to the *Buceros* (average 130 copies/genome) species. Included among the paralogous proteins in this orthogroup are the regulatory proteins tenascin-R (TNR), involved in the focal adhesion pathway, and the phosphoinositide 3-kinase (PI3K) regulatory subunit, which is involved in both the focal adhesion pathway and signalling pathways regulating pluripotency of stem cells. Gene duplications of regulatory proteins can affect gene expression and regulation in cells based on the evolutionary trajectories followed by these duplicates (Gera *et al*., 2022; Hallin and Landry, 2019). In eukaryotes, it has been observed that paralogous copies of regulatory proteins such as protein kinases can acquire new functions that are distinct from the original function through neo-functionalisation (Gao *et al*., 2015; Gera *et al*., 2022). Conversely, these proteins may undergo sub-functionalisation, where the duplicated protein specializes in managing a subset of the original targets, thereby refining their regulatory function (Gao *et al*., 2015; Gera *et al*., 2022). This dual process of neo-functionalisation and sub-functionalisation allows for the expansion and diversification of regulatory networks within cells (Gera *et al*., 2022)

3.4.6 Comparative analysis of the hornbill mitochondrial genomes

The mitochondrial genomes of the compared hornbill taxa are on average 17,157 bp in size, 373 bp larger than that of the chicken (Table 3.3). The mitochondrial genomes of the *Buceros* species are further slightly larger (on average 193 bp larger) than those of the *Bucorvus* species. The mtDNA base composition of the hornbill species and the chicken was found to be biased towards cysteine (33.66%) and adenine (31.04%), as opposed to tyrosine (22.39%) and guanine

(12.94%). This observation aligns with previous findings in the mtDNA of 61 extant hornbill species (Gonzalez *et al*., 2013).

Table 3.3: Annotation statistics of the mitochondrial genomes of the hornbill species and the chicken

	Southern Ground Hornbill	**Abyssinian Ground Hornbill**	**Great Hornbill**	**Rhinoceros Hornbill**	**Chicken**
Total size (bp)	17,061	17,061	17,117	17,390	16,784
Nucleotide frequency (%)					
Adenine (A)	30.34	30.53	30.07	30.13	30.18
Cytosine (C)	32.29	32.75	32.40	32.30	32.52
Guanine (G)	14.42	14.26	14.06	14.02	13.55
Thymine (T)	22.56	22.46	23.48	23.55	23.75
G+C content	47.11	47.01	46.46	46.32	46.07
A+T content	52.89	52.99	53.54	53.68	53.93
Length of gene content (bp)					
Protein-coding genes	11,385	11,378	11,403	11,403	11,382
tRNAs	1,532	1,532	1,534	1,534	1,525
rRNAs	2,550	2,551	2,549	2,549	2,545
Control arm (D-loop)	553	545	551	543	901

The highest sequence identity was observed between the AGH and SGH mitochondrial genomes (97.70%), while those of the *Buceros* species shared 95.00% nucleotide identity. By contrast, the *Bucorvus* mitochondrial genomes shared only 83.53% average nucleotide identity with those for the *Buceros* species, indicative of extensive mitochondrial genome diversification subsequent to differentiation of the genera (Table 3.4). Phylogenetic analysis, with the chicken mitochondrial genome as an outgroup, confirmed the clustering of the four hornbill species into the genera *Bucorvus* and *Buceros* (Figure 3.8a). This finding is consistent with the core genome phylogeny based on 7,083 SCOs (Figure 3.8b).

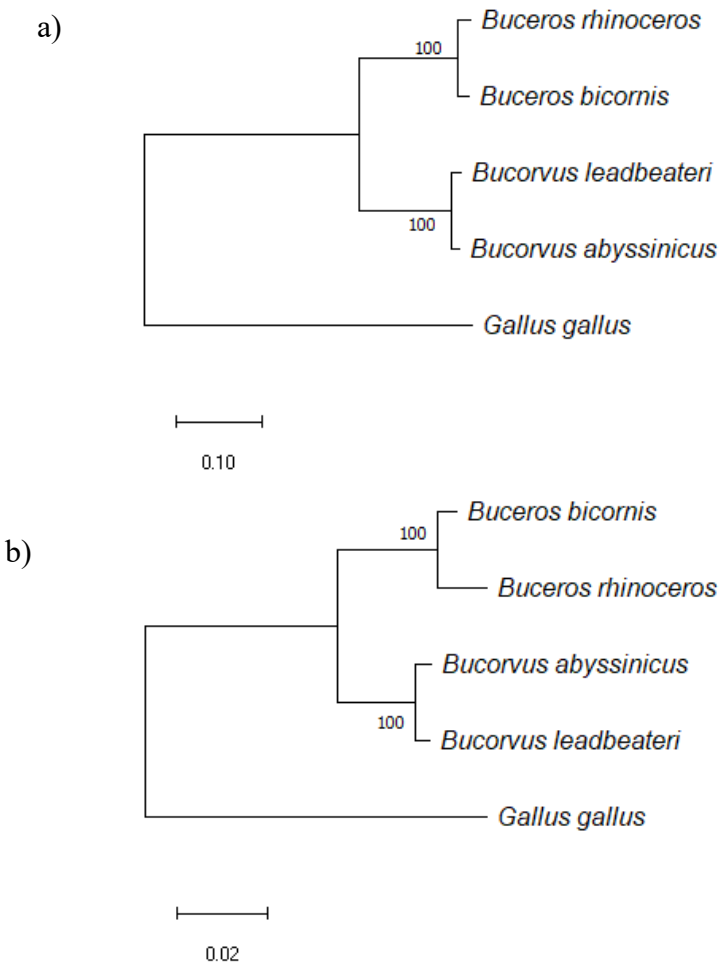


Figure 3.8: Phylogenetic placement of the four hornbills and chicken based on their mitochondrial genomes and protein homology. Phylogenetic placement of the hornbills was conducted using a) mitochondrial genomes and b) single-copy orthologues core to the hornbill and chicken genome. a) The phylogenetic tree constructed with the mitochondrial genomes was constructed with and alignment composed of 16,413 bp. A total of 16,413 sites were analysed by PhyML-SMS (Guindon *et al*., 2010) out of which 216 patterns and 11,820 sites without polymorphisms were identified. b) A total of 7,083 single-copy orthologues shared by the hornbill species hornbill species and the chicken was used to construct the phylogenetic tree. A total of 2,926,295 amino acid positions were analysed by IQ tree (Minh *et al*., 2020) during tree inference from the five sequences of which 18,538 distinct patterns, 107,486 parsimony-informative sites, 315,028 singleton sites and 2,503,781 constant sites were identified. Orthologs were refined to include the conserved regions only. A maximum likelihood tree was constructed for the mitochondrial genomes and single-copy orthologues with bootstrap support (n = 1,000). Bootstrap support greater than and equal to 100% are indicated at the nodes.

Table 3.4: Sequence similarity matrix for the mitochondrial genomes of the hornbill species and the chicken

	Abyssinian Ground Hornbill	Southern Ground Hornbill	Great Hornbill	Rhinoceros Hornbill	Chicken
Abyssinian Ground Hornbill	-	97.70	84.20	83.00	76.20
Southern Ground Hornbill	97.70	-	84.10	82.80	76.10
Great Hornbill	84.20	84.10	-	95.00	75.30
Rhinoceros Hornbill	83.00	82.80	95.00	-	74.30
Chicken	76.20	76.10	75.30	74.30	-

The mitochondrial genomes of the four hornbills and the chicken code for thirteen distinct proteins functioning in mitochondrial respiration. This includes the nicotinamide adenine dinucleotide (NAD) dehydrogenase subunits 1-6 and 4L, ATP synthases (ATP) synthase F0 subunits 6 and 8, cytochrome c oxidase (cox) subunits 1-3 and cytochrome *b* (cob). Furthermore, twenty-two tRNAs, two rRNAs and a control region with a displacement loop (D-loop) structure were identified on the mtDNA of all hornbills and the chicken. The gene order of the proteins and non-coding elements was the same in all compared taxa. The total length of the tRNA and rRNA between the hornbills was on average 1,533 bp and 2,550 bp, respectively. The total length of the protein-coding genes of the *Buceros* species was on average 22 bp and 21 bp larger than that of the *Bucorvus* species and the chicken, respectively (Table 3.3).

The mitochondrial genomes were mapped against the SGH mtDNA as a reference (Figure 3.9**)**. This highlighted the high sequence conservation between the two *Bucorvus* species and less so for the two *Buceros* species and the chicken mitochondrial genomes. In particular, low sequence conservation was observed in the control region (Figure 3.9). The control region, located between the glutamine (*trnE*) and phenylalanine (*trnF*) tRNAs on the heavy strand (Figure 3.9), is on average 549 bp in size for the hornbill species. By contrast, the control region of the chicken mitochondrial genome is on average 37.11% (324 bp) larger than that of the hornbill species. The G+C content of the control region in the *Bucorvus* species was 47.54%, this is 2.5% larger than that noted in *Buceros* species (46.35%). In comparison to the chicken which had a G+C content of 40.50% in the control region, the G+C content of the hornbills was on average 13.73% greater.

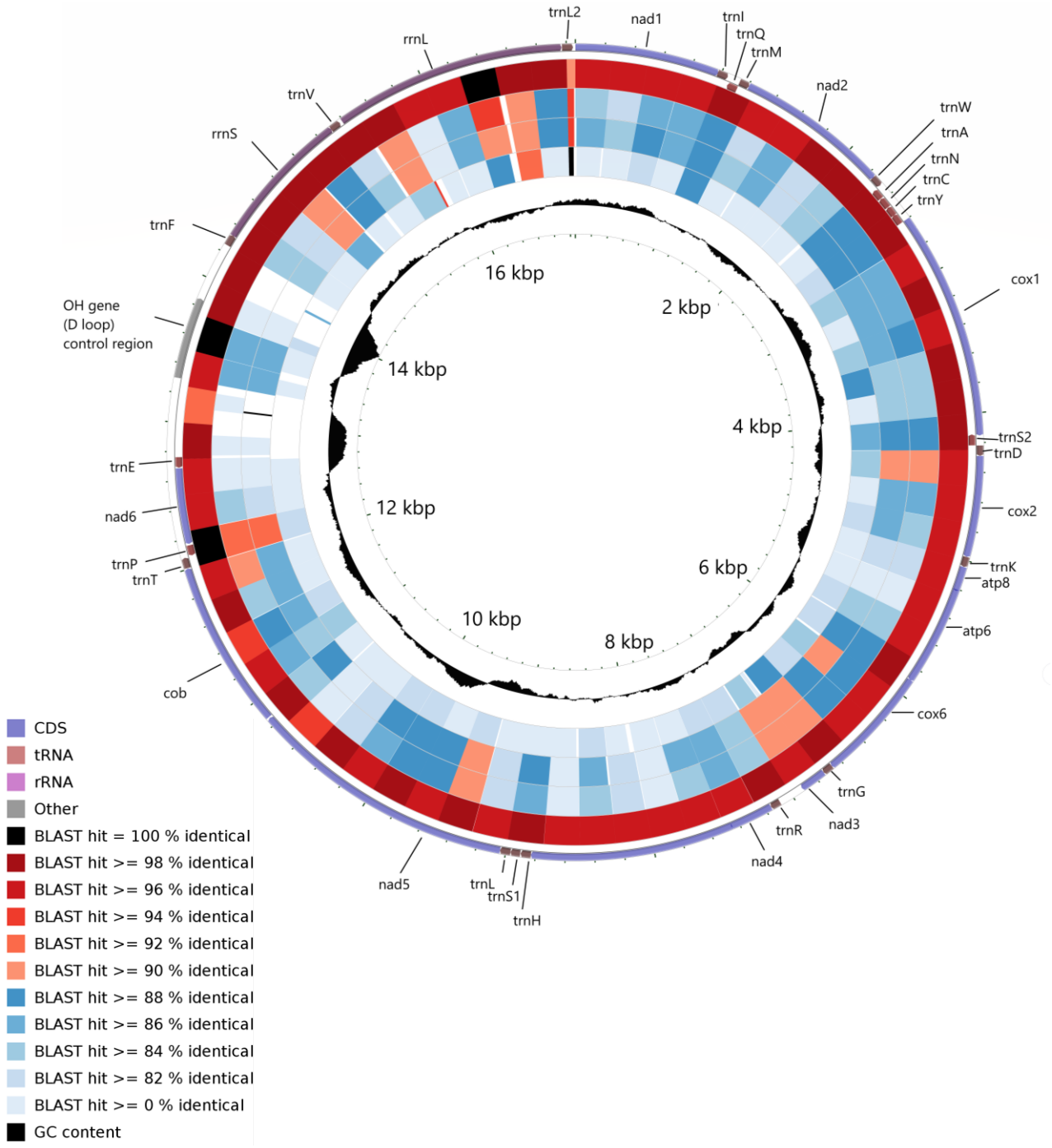


Figure 3.9: The mitochondrial genome map of the hornbill species and the chicken mapped against the SGH as a reference. The mitochondrial genome of all species was composed of thirteen proteins functioning in mitochondrial respiration, twenty-two tRNAs, two rRNAs and one control region with a displacement loop (D-loop) structure. The outermost circle represents the gene order, followed by the map of the AGH, GH, RH and lastly the chicken mtDNAs, respectively. Blastn was performed against the other mitochondrial genomes in a 250 bp window with default parameters and used to map regions of homology using CGview (Stothard and Wishart, 2005).

The control region size observed in the hornbill species in this study was significantly smaller than what has been previously reported (Chen *et al*., 2018; Delport *et al*., 2002; Lan *et al*., 2019). For example, the size of the control region in the SGH has been documented as 1,731 bp, which is 68.05% (1,178 bp) larger than that annotated in this study. Moreover, the size of the control region in hornbills, on average, has been found to be 500 bp larger than in other avian species, such as the chicken (Delport *et al*., 2002). These findings could indicate potential annotation errors due to biases in the annotation tool, artifacts, or errors in reference sequences in databases upon which these annotations are conducted (Bernt *et al*., 2013; Prada and Boore, 2019). Thus, cross-validation using multiple annotation tools is required to confirm these results. This approach would help minimize the impact of tool-specific biases and provide a more accurate comparison of the mtDNA between the hornbill species.

3.5 Discussion

3.5.1 The *Bucorvus* species exhibit greater genome sizes and genome complexity

A global analysis of the genomes of two *Bucorvus* and two *Buceros* species showed that, on average, those of the former genus are 5.58% larger than those of the latter and code for 30.31% more proteins. Avian species exhibit relatively little interspecific variations in genome size compared to other taxa (Kapusta *et al*., 2017). However, variations in genome size have been linked to factors such as metabolic rates, body size and environmental factors such as biogeographical range and demographics (Gregory *et al*., 2009; Ji *et al*., 2022; Ji and DeWoody, 2017). Recombination rates, generation time and the activity of transposable elements (TEs) (Kapusta *et al*., 2017; Nam and Ellegren, 2012) have also been identified to influence genomic diversity and consequently genome size in avian species (Gregory *et al*., 2009; Grzywacz and Skórka, 2021). The *Bucorvus* species have a substantively larger body size and lead a mainly terrestrial lifestyle in comparison to their arboreal relatives in the genus *Buceros* (Kemp and Kemp, 1980; Viseshakul *et al*., 2011). Furthermore, while the *Bucorvus* species are distributed in the African savannah leading a faunivorous lifestyle, members of the genus *Buceros* are located in the tropical forests of Asia and mainly lead a frugivorous lifestyle (Kemp and Kemp, 1980; Viseshakul *et al*., 2011). These differences in lifestyle, geographical distribution and diet could potentially contribute to the observed genetic variations between these two genera.

Compared to the sister genus, the *Bucorvus* species also exhibited a greater abundance of introns, tRNAs, rRNAs and repeat content. Non-coding elements such as tRNAs and rRNAs play a major role in the coordination of protein synthesis and regulation of gene expression. Thus, their presence and abundance within the genome could affect protein expression rates and transcriptional efficiency (Alberts *et al*., 2002b; Ottenburghs *et al*., 2021; Torrent *et al*., 2018). Repetitive elements have also been linked to regulating gene expression and further play a role in genome organisation and chromosomal instability (Benham *et al*., 2024; Feschotte, 2008; Oliveira *et al*., 2017). These elements create genetic diversity between species via recombination events (Oliveira *et al*., 2017). Thus, the variations in genome size and composition between the *Bucorvus* and *Buceros* genera could represent a multifaceted response aimed at increasing genomic diversity through the diversification of protein functions and regulatory elements. This increased genomic diversity may enhance their ability to adapt, survive and thrive in their respective environments (Blommaert, 2020).

The differences in genome sizes and structures between the two genera should, however, be interpreted cautiously due to the potential impacts of genome sequencing and assembly quality on annotated results (Hu *et al*., 2020). While the SGH, AGH, and GH genomes showed high BUSCO completeness scores (> 96%) with few fragmented genes, the RH genome scored lower (~ 89%) with numerous fragmented genes, possibly due to genome complexity or sequencing, assembly and annotation errors (Vaattovaara *et al*., 2019).

3.5.2 The *Bucorvus* species genomes code for more genus-specific proteins

Pan-genome analysis of the protein datasets encoded on the genomes of four hornbill species revealed extensive differences in the protein profiles of members of the two genera. As such, 33.82% and 28.51% of the overall orthogroups across the four taxa are specific to the genus *Bucorvus* and *Buceros*, respectively. Extensive differences in the proteomes of the individual species in each genus were also observed. For example, 7.71% and 8.94% of the orthogroups were unique to *B. abyssinicus* and *B. leadbeateri*, respectively. Amongst the orthogroups specific to the genus *Buceros* and *Bucorvus*, 96.23% and 93.70% were comprised of single-copy orthologues, with the remainder being comprised of proteins with paralogous copies. Orthologues proteins arise via speciation while paralogous proteins arise through gene duplications (David *et al*., 2020). Thus the variations noted in the *Bucorvus* and *Buceros* species could reflect differences in genetic diversity and different rates of gene gain and loss between the two genera throughout evolutionary history (Fernández and Gabaldón, 2020).

Based on the adaptation and survival of the species in the genus *Bucorvus* and *Buceros*, gene gain may have provided new functions that helped these species adapt to their environment while gene loss could indicate adaptations to specific ecological niches where certain genes are no longer necessary (Borges *et al*., 2015; Temperley *et al*., 2008). This has been observed in a study conducted on 48 avian species where a loss of two opsin genes was noted in all representatives, likely due to the degradation of the parietal organ, while a set of fifteen opsin genes were maintained to support advanced visual systems in avian species (Borges *et al*., 2015). Furthermore, changes in the opsin genes were observed at the intra-avian level. The barn owl (*Tyto alba*), for instance, exhibited signs of pseudogenisation, a mechanism that contributes to gene loss, in an opsin gene associated with light response for nocturnal adaptation (Borges *et al*., 2015).

The divergence of *Bucorvus* has been dated to have occurred approximately 13 million years ago following the appearance of extensive woodlands and grasslands, which is their preferred ecological niche (Kemp and Kemp, 1980; Viseshakul *et al*., 2011). The noted genomic differences could have contributed to the adaptation of the *Bucorvus* species to their new environments. In comparison the *Buceros* species, inhabiting the tropical forests of Asia (Viseshakul *et al*., 2011), would have had to modify their genomic repertoire to continue thriving and surviving in this environment.

It should be noted that lower quality of genome sequences poses a major limitation in genome annotation (resulting in erroneous gene predictions and the prediction of partial and missing genes) and subsequent pan-genome analysis (Hu *et al*., 2020). The RH genome incorporated much lower numbers of protein-coding and non-coding genes compared to the other hornbills, which can substantially influence downstream analysis of the pan-genome, functional analysis of proteins and phylogenetic relationships. A study conducted on 55 avian genomes of varying quality found that annotation of low-quality assemblies resulted in lower numbers of tRNA being predicted, similar to that noted in the RH (Ottenburghs *et al*., 2021). Additionally, pan-genome analysis of the chicken genome has shown that genome sequence and assembly quality can affect pan-genome analysis and downstream prediction of genes (Li *et al*., 2022). In this study, we calculated the average of genome metrics and proteins for functional analysis across the *Buceros* species to compensate for the poor quality of the RH genome. However, it is crucial to incorporate more hornbill species genomes, preferably of an accuracy and completeness comparable to the SGH and AGH in future studies to enhance the comparative power and

validate the results. This could also shed light on the influence of various factors such as lifestyle, biology and behaviour on these genomic characteristics on a broader scale across several hornbill species.

3.5.3 The *Bucorvus* and *Buceros* species evolved different repertoires of genus-specific proteins related to metabolism and organism systems

The orthogroups from the accessory genome that were specific to the genus *Bucorvus* and those to the genus *Buceros* (which excluded the species-specific orthogroups) exhibited similarities in the abundance of proteins (< 2-fold difference) associated with twenty-three of the 46 KEGG categories, which included those under the classes of "environmental information processing" and "cellular processes". Differences were, however, noted in the abundance of proteins from the orthogroups shared by each genus associated with the classes of "metabolism" and "organism systems" (i.e. excretory system, protein digestion and absorption and thermogenesis). These variations could have crucial implications for our understanding of the functional adaptation of the *Bucorvus* species following their divergence from other hornbill species.

The *Bucorvus* species exhibited a unique set of proteins associated with lipid metabolism and oxidative phosphorylation which were absent in the *Buceros* species. Avian species exhibit a greater metabolic rate (particularly lipid and protein metabolism) and oxidative stress compared to mammals due to the energy requirements of flight (McNab, 2009). Furthermore, differences in metabolic requirements have been associated with body mass, diet, environmental conditions and physiological adaptations (Song and Beissinger, 2020). In the *Bucorvus* species, the more complex presence of lipid and energy metabolism pathways compared to the *Buceros* species might be attributed to a myriad of ecological and physiological factors post its divergence from other hornbill species. These include changes in diet (faunivory vs frugivory), environmental conditions (open woodlands vs closed tree canopies), life history (terrestrial vs arboreal) and habitat (savannah vs tropical forests) (Kemp and Kemp, 1980; McNab, 2009; Viseshakul *et al*., 2011).

A study conducted on horned larks (*Eremophila alpestris*) and house sparrows (*Passer domesticus*) revealed that horned larks had a higher metabolic rate in colder temperatures and greater metabolic flexibility due to seasonal variations in open habitats compared to house sparrows in closed habitats (Oboikovitz and Swanson, 2022). Similarly, the *Bucorvus* species could have developed higher metabolic requirements to adapt to the African savannah biomes

(Kemp and Kemp, 1980; Viseshakul *et al*., 2011). These species also exhibited a greater composition of proteins associated with thermogenesis indicating a greater need for heat generation in their environment compared to the *Buceros* species. The African savannah biomes, which are grass-dominated, have scattered wood cover that varies according to the rainfall gradients and may experience greater temporal temperature variation due to their openness compared to the tree-dominated tropical forests of Asia, which provide extensive canopy cover (Anderson *et al*., 2016; Holbrook, 2008; Huntley, 2023; Johnson *et al*., 2023).

Changes in diet could also potentially have affected the digestive and metabolic physiology of the *Bucorvus* and *Buceros* species based on the changes in carbohydrate, lipid and protein intake (Wu, 2021). A frugivorous diet is high in proteins and carbohydrates, while a faunivorous diet is rich in proteins, carbohydrates, and fats, which have the highest energy and calorie content (Wu, 2021; Wu *et al*., 2021). Avian species with a frugivorous diet have been shown to exhibit lower metabolic rates compared to those with a faunivorous diet (McNab, 2009). Furthermore, the *Bucorvus* species exhibited a greater abundance of proteins associated with pathways for protein digestion and absorption compared to the *Buceros* species. This suggests that the *Bucorvus* species could have evolved a sophisticated digestive system to maximize nutrient absorption for energy production from their faunivorous diet. In contrast, the *Buceros* species have a different set of digestive and metabolic adaptations based on their fructivorous diet.

Apart from diet, habitat and environmental conditions, the lifestyles of the *Bucorvus* and *Buceros* species may also influence their energy and metabolic requirements. The *Bucorvus* species mainly lead a terrestrial lifestyle and maintain sustained flight of up to 2 to 5 km (Kemp and Kemp, 1980). Furthermore, they have been estimated to travel an average of 7.4 km^2 per day and more during the breeding season for foraging (Zoghby *et al*., 2015). Thus, they may require more energy for activities such as running and flying short distances, hence the need for more proteins associated with thermogenesis and energy production. In contrast, *Buceros* species mainly lead an arboreal lifestyle, navigating through the upper canopies of the forest and might not require as much energy for movement, as they can glide from tree to tree (Naniwadekar *et al*., 2019). Studies have shown that passerines dominating terrestrial environments have a higher metabolic requirement compared to tropical species (Bushuev *et al*., 2018; McNab, 2009). Additionally, avian species in tropical regions exhibit lower

metabolic rate requirements as a result of a slower pace of life and lower investment in reproduction and thermogenesis compared to those in temperate regions (McKechnie, 2015).

It is important to note that the observed differences in the presence or absence of accessory proteins shared between the *Bucorvus* and *Buceros* species do not necessarily imply a complete absence of these proteins in one genus and their presence in the other. These findings are based on orthology inference between the *Bucorvus* and *Buceros* species, a process that has been identified to be influenced by gene rates (slow- or fast-evolving), heterogeneity between species and algorithmic biases (Natsidis *et al*., 2021). Errors in orthology inference may result in misleading conclusions being drawn in terms of the presence or absence of genes between species (Natsidis *et al*., 2021). Therefore, the focus should not be on the binary presence or absence of these proteins, but on understanding the role these proteins may play in the unique evolutionary trajectories of these genera and how they may have shaped the biology and lifestyle of these species (Gabaldón, 2007). This perspective allows for a more nuanced and comprehensive understanding of the complex interplay between evolution, adaptation and protein function. This understanding coupled with gene expression studies of the identified proteins in the *Bucorvus* and *Buceros* species could thus provide a pivotal basis for developing effective conservation strategies for these vulnerable species in the wild and captivity.

3.5.4 The genomes of the *Bucorvus* species gained new genes through gene duplications

The formation of new genes is an important source of evolutionary and genetic novelty in species, providing a foundation upon which mutations, genetic drift and natural selection can act (Crow and Wagner, 2006). New genes can arise through various processes including gene duplications and *de novo* gene birth from ancestrally non-genic regions of a genome (Montañés *et al*., 2023). Gene duplication is one of the most common ways in which new genes are created through unequal crossing over during meiosis and retrotransposition events (Magadum *et al*., 2013). The *Bucorvus* species exhibited over 8.70 times more paralogous proteins in the core genome compared to the *Buceros* species, suggesting a higher rate of gene duplication events in the taxa belonging to the former genus. Functional analysis of these overrepresented proteins in the *Bucorvus* and *Buceros* species revealed a strong association with multiple pathways including those functioning in the “organism systems”, “environmental information processing” and “cellular processes”. This indicates that the *Bucorvus* species could have undergone gene duplications as an adaptive response to its new environment and lifestyle post-divergence. This

provides a genetic basis for the emergence of new functions and traits via neo-functionalisation or sub-functionalisation (Crow and Wagner, 2006), thereby contributing towards genetic diversity between the *Bucorvus* and *Buceros* species.

3.6 Conclusion

This study provides the first comparative genomic analysis of four hornbill species, across two genera, on a genome-wide scale. Our study revealed differences not only in genome size between these two genera, but also in the structural and functional composition of their genomes. These differences could have contributed to the different evolutionary trajectories of taxa in the two genera, particularly following the divergence of the *Bucorvus* species to the African savannah biome. Functional analysis of the accessory proteins specific to the genus *Bucorvus* and *Buceros* has further provided insight into the genetic basis underlying the evolution and adaptation of these species to their respective environments. This lays a solid foundation for further investigations into the implications of key genes involved in adaptation using gene expression studies. Moreover, this research serves as a crucial stepping-stone for large-scale comparative genomic analysis among hornbill species as more genomes become available. It is noteworthy that mitochondrial genome sequences for 61 extant hornbill species are available. Given that the family *Bucorvidae* currently constitutes only two extant species, it stands out as one of the most complete avian families with whole genomes for all representatives. This achievement highlights the significance of the *Bucorvidae* family in avian genomics and emphasizes the importance of these findings in the broader context of avian evolutionary studies. Ultimately, this research will help bridge the knowledge gap in genomic research on these birds and support the development of effective conservation strategies for their preservation.

3.7 References

Al Arab, M., Höner zu Siederdissen, C., Tout, K., Sahyoun, A. H., Stadler, P. F., and Bernt, M. (2017). Accurate annotation of protein-coding genes in mitochondrial genomes. *Molecular Phylogenetics and Evolution*, *106*, 209–216. https://doi.org/10.1016/j.ympev.2016.09.024

Albers, M., Diment, A., Muraru, M., Russell, C. S., and Beggs, J. D. (2003). Identification and characterization of Prp45p and Prp46p, essential pre-mRNA splicing factors. *RNA*, *9*(1), 138–150. https://doi.org/10.1261/rna.2119903

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002a). Eggs. In *Molecular Biology of the Cell. 4th edition*. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26842/

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002b). From RNA to Protein. In *Molecular Biology of the Cell. 4th edition*. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26829/

Anderson, T. M., White, S., Davis, B., Erhardt, R., Palmer, M., Swanson, A., Kosmala, M., and Packer, C. (2016). The spatial distribution of African savannah herbivores: Species associations and habitat occupancy in a landscape context. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1703), 20150314. https://doi.org/10.1098/rstb.2015.0314

Araújo, P. M., Viegas, I., Rocha, A. D., Villegas, A., Jones, J. G., Mendonça, L., Ramos, J. A., Masero, J. A., and Alves, J. A. (2019). Understanding how birds rebuild fat stores during migration: Insights from an experimental study. *Scientific Reports*, *9*, 10065. https://doi.org/10.1038/s41598-019-46487-z

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene Ontology: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Aslam, M., and Ladilov, Y. (2022). Emerging Role of cAMP/AMPK Signaling. *Cells*, *11*(2), Article 2. https://doi.org/10.3390/cells11020308

Ayadi, L., Callebaut, I., Saguez, C., Villa, T., Mornon, J. P., and Banroques, J. (1998). Functional and structural characterization of the Prp3 binding domain of the yeast Prp4 splicing factor1. *Journal of Molecular Biology*, *284*(3), 673–687. https://doi.org/10.1006/jmbi.1998.2183

Benham, P. M., Cicero, C., Escalona, M., Beraut, E., Fairbairn, C., Marimuthu, M. P. A., Nguyen, O., Sahasrabudhe, R., King, B. L., Thomas, W. K., Kovach, A. I., Nachman, M. W., and Bowie, R. C. K. (2024). Remarkably High Repeat Content in the Genomes of Sparrows: The Importance of Genome Assembly Completeness for Transposable Element Discovery. *Genome Biology and Evolution*, *16*(4), evae067. https://doi.org/10.1093/gbe/evae067

Berg, M. D., and Brandl, C. J. (2021). Transfer RNAs: Diversity in form and function. *RNA Biology*, *18*(3), 316–339. https://doi.org/10.1080/15476286.2020.1809197

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., Pütz, J., Middendorf, M., and Stadler, P. F. (2013). MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Bhutia, Y. D., and Ganapathy, V. (2018). Chapter 47—Protein Digestion and Absorption. In H. M. Said (Ed.), *Physiology of the Gastrointestinal Tract (Sixth Edition)* (pp. 1063–1086). Academic Press. https://doi.org/10.1016/B978-0-12-809954-4.00047-5

Blommaert, J. (2020). Genome size evolution: Towards new model systems for old questions. *Proceedings of the Royal Society B: Biological Sciences*, *287*(1933), 20201441. https://doi.org/10.1098/rspb.2020.1441

Borges, R., Khan, I., Johnson, W. E., Gilbert, M. T. P., Zhang, G., Jarvis, E. D., O'Brien, S. J., and Antunes, A. (2015). Gene loss, adaptive evolution and the co-evolution of plumage coloration genes with opsins in birds. *BMC Genomics*, *16*(1), 751. https://doi.org/10.1186/s12864-015-1924-3

Braun, E. J., and Sweazea, K. L. (2008). Glucose regulation in birds. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *151*(1), 1–9. https://doi.org/10.1016/j.cbpb.2008.05.007

Bravo, G. A., Schmitt, C. J., and Edwards, S. V. (2021). What Have We Learned from the First 500 Avian Genomes? *Annual Review of Ecology, Evolution, and Systematics*, *52*(1), 611–639. https://doi.org/10.1146/annurev-ecolsys-012121-085928

Bushuev, A., Tolstenkov, O., Zubkova, E., Solovyeva, E., and Kerimov, A. (2018). Basal metabolic rate in free-living tropical birds: The influence of phylogenetic, behavioral, and ecological factors. *Current Zoology*, *64*(1), 33–43. https://doi.org/10.1093/cz/zox018

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), 5825–5829. https://doi.org/10.1093/molbev/msab293

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): An expert resource for Glycogenomics. *Nucleic Acids Research*, *37*(Database issue), D233–D238. https://doi.org/10.1093/nar/gkn663

Chan, P. P., Lin, B. Y., Mak, A. J., and Lowe, T. M. (2021). tRNAscan-SE 2.0: Improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research*, *49*(16), 9077–9096. https://doi.org/10.1093/nar/gkab688

Chandel, N. S. (2021). Lipid Metabolism. *Cold Spring Harbor Perspectives in Biology*, *13*(9), a040576. https://doi.org/10.1101/cshperspect.a040576

Chen, Y., Li, C., Yan, H., Xiao, H., and Chen, S. (2018). The complete mitochondrial genome sequence of *Buceros* bicornis (Bucerotiformes: Bucerotidae). *Conservation Genetics Resources*, *10*(3), 287–290. https://doi.org/10.1007/s12686-017-0803-4

Cheng, Y., Miller, M. J., Zhang, D., Song, G., Jia, C., Qu, Y., and Lei, F. (2020). Comparative Genomics Reveals Evolution of a Beak Morphology Locus in a High-Altitude Songbird.

Molecular Biology and Evolution, *37*(10), 2983–2988. https://doi.org/10.1093/molbev/msaa157

Cooper, G. M. (2000). Pathways of Intracellular Signal Transduction. In *The Cell: A Molecular Approach. 2nd edition* (Second Edition). Sinauer Associates. https://www.ncbi.nlm.nih.gov/books/NBK9870/

Corioni, M., Antih, N., Tanackovic, G., Zavolan, M., and Krämer, A. (2011). Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Research*, *39*(5), 1868–1879. https://doi.org/10.1093/nar/gkq1042

Crow, K. D., and Wagner, G. P. (2006). What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, *23*(5), 887–892. https://doi.org/10.1093/molbev/msj083

Dai, Y., Tang, H., and Pang, S. (2021). The Crucial Roles of Phospholipids in Aging and Lifespan Regulation. *Frontiers in Physiology*, *12*, 775648. https://doi.org/10.3389/fphys.2021.775648

David, K. T., Oaks, J. R., and Halanych, K. M. (2020). Patterns of gene evolution following duplications and speciations in vertebrates. *PeerJ*, *8*, e8813. https://doi.org/10.7717/peerj.8813

Delport, W., Ferguson, J. W. H., and Bloomer, P. (2002). Characterization and evolution of the mitochondrial DNA control region in hornbills (Bucerotiformes). *Journal of Molecular Evolution*, *54*(6), 794–806. https://doi.org/10.1007/s00239-001-0083-0

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., Taly, J. F., and Notredame, C. (2011). T-Coffee: A web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Research*, *39*(Web Server issue), W13–W17. https://doi.org/10.1093/nar/gkr245

Donath, A., Jühling, F., Al-Arab, M., Bernhart, S. H., Reinhardt, F., Stadler, P. F., Middendorf, M., and Bernt, M. (2019). Improved annotation of protein-coding genes boundaries in

metazoan mitochondrial genomes. *Nucleic Acids Research*, *47*(20), 10543–10552. https://doi.org/10.1093/nar/gkz833

Doudna, J. A., and Rath, V. L. (2002). Structure and Function of the Eukaryotic Ribosome: The Next Frontier. *Cell*, *109*(2), 153–156. https://doi.org/10.1016/S0092-8674(02)00725-0

DuBose, T. D. (1990). Reclamation of filtered bicarbonate. *Kidney International*, *38*(4), 584–589. https://doi.org/10.1038/ki.1990.246

Earnshaw, C. H., and Warren, R. B. (2023). Interleukin (IL)-12/23 and IL-23 Inhibitors. In *Handbook of Systemic Drug Treatment in Dermatology* (3rd ed.). CRC Press.

Eliason, C. M., Mellenthin, L. E., Hains, T., McCullough, J. M., Pirro, S., Andersen, M. J., and Hackett, S. J. (2023). Genomic signatures of convergent shifts to plunge-diving behavior in birds. *Communications Biology*, *6*(1), Article 1. https://doi.org/10.1038/s42003-023-05359-z

EMBOSS. (2002). *EMBOSS: seqret manual*. Seqret. https://www.bioinformatics.nl/cgi-bin/emboss/help/seqret

Emms, D. M., and Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), Article 1. https://doi.org/10.1186/s13059-015-0721-2

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Feng, S., Stiller, J., Deng, Y., Armstrong, J., Fang, Q., Reeve, A. H., Xie, D., Chen, G., Guo, C., Faircloth, B. C., Petersen, B., Wang, Z., Zhou, Q., Diekhans, M., Chen, W., Andreu-Sánchez, S., Margaryan, A., Howard, J. T., Parent, C., … Zhang, G. (2020). Dense sampling of bird diversity increases power of comparative genomics. *Nature*, *587*(7833), 252–257. https://doi.org/10.1038/s41586-020-2873-9

Fernández, R., and Gabaldón, T. (2020). Gene gain and loss across the Metazoa Tree of Life. *Nature Ecology and Evolution*, *4*(4), 524–533. https://doi.org/10.1038/s41559-019-1069-x

Feschotte, C. (2008). The contribution of transposable elements to the evolution of regulatory networks. *Nature Reviews. Genetics*, *9*(5), 397–405. https://doi.org/10.1038/nrg2337

Figueroa, J. D., and Hayman, M. J. (2004). The human Ski-interacting protein functionally substitutes for the yeast *PRP45* gene. *Biochemical and Biophysical Research Communications*, *319*(4), 1105–1109. https://doi.org/10.1016/j.bbrc.2004.05.096

Fleming, P. A., and Nicolson, S. W. (2003). Osmoregulation in an avian nectarivore, the whitebellied sunbird Nectarinia talatala: Response to extremes of diet concentration. *Journal of Experimental Biology*, *206*(11), 1845–1854. https://doi.org/10.1242/jeb.00351

Gabaldón, T. (2007). Evolution of proteins and proteomes: A phylogenetics approach. *Evolutionary Bioinformatics Online*, *1*, 51–61.

Gabaldón, T., and Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews. Genetics*, *14*(5), 360–366. https://doi.org/10.1038/nrg3456

Gao, D., Ko, D. C., Tian, X., Yang, G., and Wang, L. (2015). Expression Divergence of Duplicate Genes in the Protein Kinase Superfamily in Pacific Oyster. *Evolutionary Bioinformatics*, *11s1*, EBO.S30230. https://doi.org/10.4137/EBO.S30230

Gene Ontology Consortium, Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., Feuermann, M., Gaudet, P., Harris, N. L., Hill, D. P., Lee, R., Mi, H., Moxon, S., Mungall, C. J., Muruganugan, A., Mushayahama, T., Sternberg, P. W., Thomas, P. D., … Westerfield, M. (2023). The Gene Ontology knowledgebase in 2023. *Genetics*, *224*(1), iyad031. https://doi.org/10.1093/genetics/iyad031

Gera, T., Jonas, F., More, R., and Barkai, N. (2022). Evolution of binding preferences among whole-genome duplicated transcription factors. *eLife*, *11*, e73225. https://doi.org/10.7554/eLife.73225

Goldstein, D. L. (2006). Regulation of the avian kidney by arginine vasotocin. *General and Comparative Endocrinology*, *147*(1), 78–84. https://doi.org/10.1016/j.ygcen.2005.09.018

Goldstrohm, A. C., Albrecht, T. R., Suñé, C., Bedford, M. T., and Garcia-Blanco, M. A. (2001). The Transcription Elongation Factor CA150 Interacts with RNA Polymerase II and the Pre-mRNA Splicing Factor SF1. *Molecular and Cellular Biology*. https://doi.org/10.1128/MCB.21.22.7617-7628.2001

Gonzalez, J. C. T., Sheldon, B. C., Collar, N. J., and Tobias, J. A. (2013). A comprehensive molecular phylogeny for the hornbills (Aves: Bucerotidae). *Molecular Phylogenetics and Evolution*, *67*(2), 468–483. https://doi.org/10.1016/j.ympev.2013.02.012

Gregory, T. R., Andrews, C. B., McGuire, J. A., and Witt, C. C. (2009). The smallest avian genomes are found in hummingbirds. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1674), 3753–3757. https://doi.org/10.1098/rspb.2009.1004

Grzywacz, B., and Skórka, P. (2021). Genome size versus geographic range size in birds. *PeerJ*, *9*, e10868. https://doi.org/10.7717/peerj.10868

Guglielmo, C. G., Gerson, A. R., Price, E. R., and Hays, Q. R. (2017). The effects of dietary macronutrients on flight ability, energetics, and fuel metabolism of yellow-rumped warblers Setophaga coronata. *Journal of Avian Biology*, *48*(1), 133–148. https://doi.org/10.1111/jav.01351

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Gyamfi, D., Ofori Awuah, E., and Owusu, S. (2019). Chapter 2 - Lipid Metabolism: An Overview. In V. B. Patel (Ed.), *The Molecular Nutrition of Fats* (pp. 17–32). Academic Press. https://doi.org/10.1016/B978-0-12-811297-7.00002-0

Hallin, J., and Landry, C. R. (2019). Regulation plays a multifaceted role in the retention of gene duplicates. *PLoS Biology*, *17*(11), e3000519. https://doi.org/10.1371/journal.pbio.3000519

Hardie, D. G. (2014). AMPK--sensing energy while talking to other signaling pathways. *Cell Metabolism*, *20*(6), 939–952. https://doi.org/10.1016/j.cmet.2014.09.013

Harrington, L. E., Hatton, R. D., Mangan, P. R., Turner, H., Murphy, T. L., Murphy, K. M., and Weaver, C. T. (2005). Interleukin 17–producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nature Immunology*, *6*(11), 1123–1132. https://doi.org/10.1038/ni1254

Harrison, D. A. (2012). The JAK/STAT Pathway. *Cold Spring Harbor Perspectives in Biology*, *4*(3), a011205. https://doi.org/10.1101/cshperspect.a011205

Harvey, M. G., Bravo, G. A., Claramunt, S., Cuervo, A. M., Derryberry, G. E., Battilana, J., Seeholzer, G. F., McKay, J. S., O'Meara, B. C., Faircloth, B. C., Edwards, S. V., Pérez-Emán, J., Moyle, R. G., Sheldon, F. H., Aleixo, A., Smith, B. T., Chesser, R. T., Silveira, L. F., Cracraft, J., … Derryberry, E. P. (2020). The evolution of a tropical biodiversity hotspot. *Science*, *370*(6522), 1343–1348. https://doi.org/10.1126/science.aaz6970

Hegemann, A., Fudickar, A. M., and Nilsson, J. Å. (2019). A physiological perspective on the ecology and evolution of partial migration. *Journal of Ornithology*, *160*(3), 893–905. https://doi.org/10.1007/s10336-019-01648-9

Hegemann, A., Matson, K. D., Both, C., and Tieleman, B. I. (2012). Immune function in a free-living bird varies over the annual cycle, but seasonal patterns differ between years. *Oecologia*, *170*(3), 605–618. https://doi.org/10.1007/s00442-012-2339-3

Hermier, D. (1997). Lipoprotein Metabolism and Fattening in Poultry1. *The Journal of Nutrition*, *127*(5), 805S-808S. https://doi.org/10.1093/jn/127.5.805S

Hillier, L. W., Miller, W., Birney, E., Warren, W., Hardison, R. C., Ponting, C. P., Bork, P., Burt, D. W., Groenen, M. A. M., Delany, M. E., Dodgson, J. B., Chinwalla, A. T., Cliften, P. F., Clifton, S. W., Delehaunty, K. D., Fronick, C., Fulton, R. S., Graves, T. A., Kremitzki, C., … Project management: (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, *432*(7018), Article 7018. https://doi.org/10.1038/nature03154

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hoang Do, O., and Thorn, P. (2015). Insulin secretion from beta cells within intact islets: Location matters. *Clinical and Experimental Pharmacology and Physiology*, *42*(4), 406–414. https://doi.org/10.1111/1440-1681.12368

Hohtola, E. (2012). Thermoregulatory Adaptations to Starvation in Birds. In M. D. McCue (Ed.), *Comparative Physiology of Fasting, Starvation, and Food Limitation* (pp. 155–170). Springer. https://doi.org/10.1007/978-3-642-29056-5_10

Holbrook, K. (2008). The Ecology and Conservation of Asian Hornbills: Farmers of the Forest. *Condor*, *110*, 810–811. https://doi.org/10.1525/cond.2008.8657

Horrocks, N. P. C., Hegemann, A., Ostrowski, S., Ndithia, H., Shobrak, M., Williams, J. B., Matson, K. D., and Tieleman, B. I. (2015). Environmental proxies of antigen exposure explain variation in immune investment better than indices of pace of life. *Oecologia*, *177*(1), 281–290. https://doi.org/10.1007/s00442-014-3136-y

Hu, X., Wang, Y., Sheikhahmadi, A., Li, X., Buyse, J., Lin, H., and Song, Z. (2019). Effects of dietary energy level on appetite and central adenosine monophosphate-activated protein kinase (AMPK) in broilers. *Journal of Animal Science*, *97*(11), 4488–4495. https://doi.org/10.1093/jas/skz312

Hu, Y., Huang, J., Chen, C., Wang, Y., Hao, Z., Chen, T., Wang, J., and Li, J. (2022). Strategies of Macrophages to Maintain Bone Homeostasis and Promote Bone Repair: A Narrative Review. *Journal of Functional Biomaterials*, *14*(1), 18. https://doi.org/10.3390/jfb14010018

Hu, Z., Wei, C., and Li, Z. (2020). Computational Strategies for Eukaryotic Pangenome Analyses. In H. Tettelin and D. Medini (Eds.), *The Pangenome: Diversity, Dynamics and Evolution of Genomes* (pp. 293–307). Springer International Publishing. https://doi.org/10.1007/978-3-030-38281-0_13

Hughes, M. R. (2003). Regulation of salt gland, gut and kidney interactions. *Comparative Biochemistry and Physiology. Part A, Molecular and Integrative Physiology*, *136*(3), 507–524. https://doi.org/10.1016/j.cbpb.2003.09.005

Huntley, B. J. (2023). The Mesic Savanna Biome. In B. J. Huntley (Ed.), *Ecology of Angola: Terrestrial Biomes and Ecoregions* (pp. 313–338). Springer International Publishing. https://doi.org/10.1007/978-3-031-18923-4_14

Iridian Genomes. (2023). *IRIDIAN GENOMES*. https://www.iridiangenomes.com/contact.html

Isailovic, N., Daigo, K., Mantovani, A., and Selmi, C. (2015). Interleukin-17 and innate immunity in infections and chronic inflammation. *Journal of Autoimmunity*, *60*, 1–11. https://doi.org/10.1016/j.jaut.2015.04.006

Isaksson, C., Andersson, M. N., Nord, A., von Post, M., and Wang, H. L. (2017). Species-Dependent Effects of the Urban Environment on Fatty Acid Composition and Oxidative Stress in Birds. *Frontiers in Ecology and Evolution*, *5*. https://doi.org/10.3389/fevo.2017.00044

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz, B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L., Narula, N., Liu, L., … Zhang, G. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, *346*(6215), 1320–1331. https://doi.org/10.1126/science.1253451

Ji, J., Tao, Y., Zhang, X., Pan, J., Zhu, X., Wang, H., Du, P., Zhu, Y., Huang, Y., and Chen, W. (2020). Dynamic changes of blood glucose, serum biochemical parameters and gene expression in response to exogenous insulin in Arbor Acres broilers and Silky fowls. *Scientific Reports*, *10*(1), 6697. https://doi.org/10.1038/s41598-020-63549-9

Ji, Y., and DeWoody, J. A. (2017). Relationships Among Powered Flight, Metabolic Rate, Body Mass, Genome Size, and the Retrotransposon Complement of Volant Birds. *Evolutionary Biology*, *44*(2), 261–272. https://doi.org/10.1007/s11692-016-9405-4

Ji, Y., Feng, S., Wu, L., Fang, Q., Brüniche-Olsen, A., DeWoody, J. A., Cheng, Y., Zhang, D., Hao, Y., Song, G., Qu, Y., Suh, A., Zhang, G., Hackett, S. J., and Lei, F. (2022). Orthologous microsatellites, transposable elements, and DNA deletions correlate with generation time and body mass in neoavian birds. *Science Advances*, *8*(35), eabo0099. https://doi.org/10.1126/sciadv.abo0099

Johnson, D., Liu, X., and Burslem, D. F. R. P. (2023). Symbiotic control of canopy dominance in subtropical and tropical forests. *Trends in Plant Science*, *28*(9), 995–1003. https://doi.org/10.1016/j.tplants.2023.03.027

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S., and Madden, T. L. (2008). NCBI BLAST: A better web interface. *Nucleic Acids Research*, *36*(suppl_2), W5–W9. https://doi.org/10.1093/nar/gkn201

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., von Haeseler, A., and Jermiin, L. S. (2017). ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*(suppl_1), D480–D484. https://doi.org/10.1093/nar/gkm882

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30.

Kanehisa, M., and Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science: A Publication of the Protein Society*, *29*(1), 28–35. https://doi.org/10.1002/pro.3711

Kanehisa, M., Sato, Y., and Kawashima, M. (2022). KEGG mapping tools for uncovering hidden features in biological data. *Protein Science: A Publication of the Protein Society*, *31*(1), 47–53. https://doi.org/10.1002/pro.4172

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), 164–185. https://doi.org/10.1111/nyas.13295

Kapusta, A., Suh, A., and Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, *114*(8), E1460–E1469. https://doi.org/10.1073/pnas.1616702114

Katoh, K., Rozewicki, J., and Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160–1166. https://doi.org/10.1093/bib/bbx108

Kemp, A. C., and Kemp, M. I. (1980). The biology of the Southern Ground Hornbill *Bucorvus* leadbeateri (Vigors) (Aves: Bucerotidae). *Annals of the Transvaal Museum*, *32*(4), 65–100. https://doi.org/10.10520/AJA00411752_1057

Kim, J. H., and Kim, N. (2016). Signaling Pathways in Osteoclast Differentiation. *Chonnam Medical Journal*, *52*(1), 12–17. https://doi.org/10.4068/cmj.2016.52.1.12

Kondo, Y., Oubridge, C., van Roon, A. M. M., and Nagai, K. (2015). Crystal structure of human U1 snRNP, a small nuclear ribonucleoprotein particle, reveals the mechanism of 5′ splice site recognition. *eLife*, *4*, e04986. https://doi.org/10.7554/eLife.04986

Koonin, E. V. (2005). Orthologs, Paralogs, and Evolutionary Genomics. *Annual Review of Genetics*, *39*(1), 309–338. https://doi.org/10.1146/annurev.genet.39.073003.114725

Korn, T., Bettelli, E., Gao, W., Awasthi, A., Jäger, A., Strom, T. B., Oukka, M., and Kuchroo, V. K. (2007). IL-21 initiates an alternative pathway to induce proinflammatory TH17 cells. *Nature*, *448*(7152), 484–487. https://doi.org/10.1038/nature05970

Lamond, A. I. (1993). The spliceosome. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *15*(9), 595–603. https://doi.org/10.1002/bies.950150905

Lan, Y., Liu, M., and Cao, Y. (2019). The complete mitochondrial genome of Rhinoceros hornbill (Bucerotiformes: Bucerotidae). *Conservation Genetics Resources*, *11*(1), 75–78. https://doi.org/10.1007/s12686-017-0972-1

Lefort, V., Longueville, J. E., and Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular Biology and Evolution*, *34*(9), 2422–2424. https://doi.org/10.1093/molbev/msx149

Li, M., Sun, C., Xu, N., Bian, P., Tian, X., Wang, X., Wang, Y., Jia, X., Heller, R., Wang, M., Wang, F., Dai, X., Luo, R., Guo, Y., Wang, X., Yang, P., Hu, D., Liu, Z., Fu, W., … Yang, N. (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), msac066. https://doi.org/10.1093/molbev/msac066

Locati, M. D., Pagano, J. F. B., Girard, G., Ensink, W. A., Olst, M. van, Leeuwen, S. van, Nehrdich, U., Spaink, H. P., Rauwerda, H., Jonker, M. J., Dekker, R. J., and Breit, T. M. (2017). Expression of distinct maternal and somatic 5.8S, 18S, and 28S rRNA types during zebrafish development. *RNA*, *23*(8), 1188. https://doi.org/10.1261/rna.061515.117

Lynch, M., and Conery, J. S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. *Science*, *290*(5494), 1151–1155. https://doi.org/10.1126/science.290.5494.1151

Ma, X., Liu, B., Gong, Z., Yu, X., and Cai, J. (2021). Structural and Evolutionary Adaptation of NOD-Like Receptors in Birds. *BioMed Research International*, *2021*, 5546170. https://doi.org/10.1155/2021/5546170

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), 155–161. https://doi.org/10.1007/s12041-013-0212-8

Martin, R., Hackl, T., Hattab, G., Fischer, M. G., and Heider, D. (2021). MOSGA: Modular Open-Source Genome Annotator. *Bioinformatics*, *36*(22–23), 5514–5515. https://doi.org/10.1093/bioinformatics/btaa1003

Matera, A. G., and Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, *15*(2), 108–121. https://doi.org/10.1038/nrm3742

McGrath, C. L., Gout, J. F., Johri, P., Doak, T. G., and Lynch, M. (2014). Differential retention and divergent resolution of duplicate genes following whole-genome duplication. *Genome Research*, *24*(10), 1665–1675. https://doi.org/10.1101/gr.173740.114

McKechnie, A. E. (2015). Global variation in avian metabolic rates and the slow pace of life of tropical birds. *Functional Ecology*, *29*(3), 297–298. https://doi.org/10.1111/1365-2435.12350

McNab, B. K. (2009). Ecological factors affect the level and scaling of avian BMR. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, *152*(1), 22–45. https://doi.org/10.1016/j.cbpa.2008.08.021

Miller, A. E., Hu, P., and Barker, T. H. (2020). Feeling Things Out: Bidirectional Signaling of the Cell-ECM Interface, Implications in the Mechanobiology of Cell Spreading, Migration, Proliferation, and Differentiation. *Advanced Healthcare Materials*, *9*(8), e1901445. https://doi.org/10.1002/adhm.201901445

Mills, K. H. G. (2023). IL-17 and IL-17-producing cells in protection versus pathology. *Nature Reviews Immunology*, *23*(1), 38–54. https://doi.org/10.1038/s41577-022-00746-9

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), Article 5. https://doi.org/10.1093/molbev/msaa015

Molina, J. R., and Adjei, A. A. (2006). The Ras/Raf/MAPK Pathway. *Journal of Thoracic Oncology*, *1*(1), 7–9. https://doi.org/10.1016/S1556-0864(15)31506-9

Montañés, J. C., Huertas, M., Messeguer, X., and Albà, M. M. (2023). Evolutionary Trajectories of New Duplicated and Putative De Novo Genes. *Molecular Biology and Evolution*, *40*(5), msad098. https://doi.org/10.1093/molbev/msad098

Nam, K., and Ellegren, H. (2012). Recombination Drives Vertebrate Genome Contraction. *PLOS Genetics*, *8*(5), e1002680. https://doi.org/10.1371/journal.pgen.1002680

Naniwadekar, R., Rathore, A., Shukla, U., Chaplod, S., and Datta, A. (2019). How far do Asian forest hornbills disperse seeds? *Acta Oecologica*, *101*, 103482. https://doi.org/10.1016/j.actao.2019.103482

Natsidis, P., Kapli, P., Schiffer, P. H., and Telford, M. J. (2021). Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience*, *24*(2), 102110. https://doi.org/10.1016/j.isci.2021.102110

Neath, A. A., and Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics*, *4*(2), 199–203. https://doi.org/10.1002/wics.199

Nguyen, L. T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Nguyen, T. H. D., Galej, W. P., Bai, X., Savva, C. G., Newman, A. J., Scheres, S. H. W., and Nagai, K. (2015). The architecture of the spliceosomal U4/U6.U5 tri-snRNP. *Nature*, *523*(7558), 47–52. https://doi.org/10.1038/nature14548

Nolfi-Donegan, D., Braganza, A., and Shiva, S. (2020). Mitochondrial electron transport chain: Oxidative phosphorylation, oxidant production, and methods of measurement. *Redox Biology*, *37*, 101674. https://doi.org/10.1016/j.redox.2020.101674

Oboikovitz, P., and Swanson, D. L. (2022). Seasonal metabolic flexibility is correlated with microclimate variation in horned larks and house sparrows. *Current Zoology*, *68*(2), 199–210. https://doi.org/10.1093/cz/zoab037

Oliveira, T. D. de, Kretschmer, R., Bertocchi, N. A., Degrandi, T. M., Oliveira, E. H. C. de, Cioffi, M. de B., Garnero, A. del V., and Gunski, R. J. (2017). Genomic Organization of Repetitive DNA in Woodpeckers (Aves, Piciformes): Implications for Karyotype and ZW Sex Chromosome Differentiation. *PLOS ONE*, *12*(1), e0169987. https://doi.org/10.1371/journal.pone.0169987

Oliveros, C. H., Field, D. J., Ksepka, D. T., Barker, F. K., Aleixo, A., Andersen, M. J., Alström, P., Benz, B. W., Braun, E. L., Braun, M. J., Bravo, G. A., Brumfield, R. T., Chesser, R. T., Claramunt, S., Cracraft, J., Cuervo, A. M., Derryberry, E. P., Glenn, T. C., Harvey, M. G., … Faircloth, B. C. (2019). Earth history and the passerine superradiation. *Proceedings of the National Academy of Sciences*, *116*(16), 7916–7925. https://doi.org/10.1073/pnas.1813206116

Ottenburghs, J., Geng, K., Suh, A., and Kutter, C. (2021). Genome Size Reduction and Transposon Activity Impact tRNA Gene Diversity While Ensuring Translational

Stability in Birds. *Genome Biology and Evolution*, *13*(4), evab016. https://doi.org/10.1093/gbe/evab016

Pacheco, M. A., Battistuzzi, F. U., Lentino, M., Aguilar, R. F., Kumar, S., and Escalante, A. A. (2011). Evolution of Modern Birds Revealed by Mitogenomics: Timing the Radiation and Origin of Major Orders. *Molecular Biology and Evolution*, *28*(6), 1927–1942. https://doi.org/10.1093/molbev/msr014

Pani, P., and Bal, N. C. (2022). Avian adjustments to cold and non-shivering thermogenesis: Whats, wheres and hows. *Biological Reviews*, *97*(6), 2106–2126. https://doi.org/10.1111/brv.12885

Park, H., Li, Z., Yang, X. O., Chang, S. H., Nurieva, R., Wang, Y. H., Wang, Y., Hood, L., Zhu, Z., Tian, Q., and Dong, C. (2005). A distinct lineage of CD4 T cells regulates tissue inflammation by producing interleukin 17. *Nature Immunology*, *6*(11), 1133–1141. https://doi.org/10.1038/ni1261

Prada, C. F., and Boore, J. L. (2019). Gene annotation errors are common in the mammalian mitochondrial genomes database. *BMC Genomics*, *20*(1), 73. https://doi.org/10.1186/s12864-019-5447-1

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P., Lemmon, E. M., and Lemmon, A. R. (2015). A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature*, *526*(7574), Article 7574. https://doi.org/10.1038/nature15697

Qu, Y., Zhao, H., Han, N., Zhou, G., Song, G., Gao, B., Tian, S., Zhang, J., Zhang, R., Meng, X., Zhang, Y., Zhang, Y., Zhu, X., Wang, W., Lambert, D., Ericson, P. G. P., Subramanian, S., Yeung, C., Zhu, H., … Lei, F. (2013). Ground tit genome reveals avian adaptation to living at high altitudes in the Tibetan plateau. *Nature Communications*, *4*(1), 2071. https://doi.org/10.1038/ncomms3071

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Kim, J., Lee, C., Ko, B. J., Chaisson, M., Gedman, G. L., Cantin, L. J., Thibaud-Nissen, F., Haggerty, L., Bista, I., Smith, M., … Jarvis, E. D.

(2021). Towards complete and error-free genome assemblies of all vertebrate species. *Nature*, *592*(7856), Article 7856. https://doi.org/10.1038/s41586-021-03451-0

Roberts, A. D., and Taraska, J. W. (2023). Chapter Eight—B cell receptor (BCR) endocytosis. In I. Mani and V. Singh (Eds.), *Progress in Molecular Biology and Translational Science* (Vol. 194, pp. 159–177). Academic Press. https://doi.org/10.1016/bs.pmbts.2022.08.003

Sabat, P., Gonzalez-Vejares, S., and Maldonado, K. (2009). Diet and habitat aridity affect osmoregulatory physiology: An intraspecific field study along environmental gradients in the Rufous-collared sparrow. *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, *152*(3), 322–326. https://doi.org/10.1016/j.cbpa.2008.11.003

Sammler, S., Bleidorn, C., and Tiedemann, R. (2011). Full mitochondrial genome sequences of two endemic Philippine hornbill species (Aves: Bucerotidae) provide evidence for pervasive mitochondrial DNA recombination. *BMC Genomics*, *12*, 35. https://doi.org/10.1186/1471-2164-12-35

Scanes, C. G. (2020). Avian Physiology: Are Birds Simply Feathered Mammals? *Frontiers in Physiology*, *11*, 542466. https://doi.org/10.3389/fphys.2020.542466

Schneider, W. J. (2009). Receptor-mediated mechanisms in ovarian follicle and oocyte development. *General and Comparative Endocrinology*, *163*(1), 18–23. https://doi.org/10.1016/j.ygcen.2008.11.032

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences*, *95*(11), 5857–5864. https://doi.org/10.1073/pnas.95.11.5857

Seemann, T., and Booth, T. (2013). *Barrnap* [Computer software].

Selle, PeterH., and Liu, S. Y. (2019). The Relevance of Starch and Protein Digestive Dynamics in Poultry. *Journal of Applied Poultry Research*, *28*(3), 531–545. https://doi.org/10.3382/japr/pfy026

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), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slack, C. (2017). Ras signaling in aging and metabolic regulation. *Nutrition and Healthy Aging*, *4*(3), 195–205. https://doi.org/10.3233/NHA-160021

Smit, A. F. A., Hubley, R., and Green, P. (1996). *RepeatMasker*. RepeatMasker. http://www.repeatmasker.org

Song, S., and Beissinger, S. R. (2020). Environmental and ecological correlates of avian field metabolic rate and water flux. *Functional Ecology*, *34*(4), 811–821. https://doi.org/10.1111/1365-2435.13526

Sonnhammer, E. L., Eddy, S. R., and Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. *Proteins*, *28*(3), 405–420. https://doi.org/10.1002/(sici)1097-0134(199707)28:3<405::aid-prot10>3.0.co;2-l

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: Ab initio prediction of alternative transcripts. *Nucleic Acids Research*, *34*(suppl_2), W435–W439. https://doi.org/10.1093/nar/gkl200

Stiller, J., and Zhang, G. (2019). Comparative Phylogenomics, a Stepping Stone for Bird Biodiversity Studies. *Diversity*, *11*(7), Article 7. https://doi.org/10.3390/d11070115

Stothard, P., and Wishart, D. S. (2005). Circular genome visualization and exploration using CGView. *Bioinformatics (Oxford, England)*, *21*(4), 537–539. https://doi.org/10.1093/bioinformatics/bti054

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, *56*(4), 564–577. https://doi.org/10.1080/10635150701472164

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, *28*(1), 33–36.

Temperley, N. D., Berlin, S., Paton, I. R., Griffin, D. K., and Burt, D. W. (2008). Evolution of the chicken Toll-like receptor gene family: A story of gene gain and gene loss. *BMC Genomics*, *9*(1), 62. https://doi.org/10.1186/1471-2164-9-62

Tesmer, L. A., Lundy, S. K., Sarkar, S., and Fox, D. A. (2008). Th17 cells in human disease. *Immunological Reviews*, *223*, 87–113. https://doi.org/10.1111/j.1600-065X.2008.00628.x

Tettelin, H., Riley, D., Cattuto, C., and Medini, D. (2008). Comparative genomics: The bacterial pan-genome. *Current Opinion in Microbiology*, *11*(5), 472–477. https://doi.org/10.1016/j.mib.2008.09.006

Tokushima, Y., Takahashi, K., Sato, K., and Akiba, Y. (2005). Glucose uptake in vivo in skeletal muscles of insulin-injected chicks. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology*, *141*(1), 43–48. https://doi.org/10.1016/j.cbpc.2005.01.008

Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., and Babu, M. M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science Signaling*, *11*(546), eaat6409. https://doi.org/10.1126/scisignal.aat6409

Vaattovaara, A., Leppälä, J., Salojärvi, J., and Wrzaczek, M. (2019). High-throughput sequencing data and the impact of plant gene annotation quality. *Journal of Experimental Botany*, *70*(4), 1069–1076. https://doi.org/10.1093/jxb/ery434

Valls, P. O., and Esposito, A. (2022). Signalling dynamics, cell decisions, and homeostatic control in health and disease. *Current Opinion in Cell Biology*, *75*, None. https://doi.org/10.1016/j.ceb.2022.01.011

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, *10*(10), 725–732. https://doi.org/10.1038/nrg2600

Vertebrate Genomes Project. (2023, April 3). *Vertebrate Genomes Project*. Vertebrate Genomes Project. https://vertebrategenomesproject.org

Vianna, J. A., Fernandes, F. A. N., Frugone, M. J., Figueiró, H. V., Pertierra, L. R., Noll, D., Bi, K., Wang-Claypool, C. Y., Lowther, A., Parker, P., Le Bohec, C., Bonadonna, F., Wienecke, B., Pistorius, P., Steinfurth, A., Burridge, C. P., Dantas, G. P. M., Poulin, E., Simison, W. B., … Bowie, R. C. K. (2020). Genome-wide analyses reveal drivers of penguin diversification. *Proceedings of the National Academy of Sciences*, *117*(36), 22303–22310. https://doi.org/10.1073/pnas.2006659117

Viseshakul, N., Charoennitikul, W., Kitamura, S., Kemp, A., Thong-Aree, S., Surapunpitak, Y., Poonswad, P., and Ponglikitmongkol, M. (2011). A phylogeny of frugivorous hornbills linked to the evolution of Indian plants within Asian rainforests. *Journal of Evolutionary Biology*, *24*(7), 1533–1545. https://doi.org/10.1111/j.1420-9101.2011.02285.x

Wagner, C. A., Devuyst, O., Bourgeois, S., and Mohebbi, N. (2009). Regulated acid-base transport in the collecting duct. *Pflugers Archiv: European Journal of Physiology*, *458*(1), 137–156. https://doi.org/10.1007/s00424-009-0657-z

Wallace, I. M., O'Sullivan, O., Higgins, D. G., and Notredame, C. (2006). M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *34*(6), 1692–1699. https://doi.org/10.1093/nar/gkl091

Wang, Y., Buyse, J., Courousse, N., Tesseraud, S., Métayer-Coustard, S., Berri, C., Schallier, S., Everaert, N., and Collin, A. (2020). Effects of sex and fasting/refeeding on hepatic AMPK signaling in chickens (*Gallus gallus*). *Comparative Biochemistry and Physiology Part A: Molecular and Integrative Physiology*, *240*, 110606. https://doi.org/10.1016/j.cbpa.2019.110606

Warren, W., Fedrigo, O., Tracey, A., Mason, A. S., Formenti, G., Francesco, P., Wu, Z., Murphy, T., Schneider, V., Stiers, K., Rice, E. S., Coghill, L. M., Anthony, N., Okimoto, R., Carroll, R., Mountcastle, J., Balacco, J., Haase, B., Yang, C., … Jarvis, E. D. (2023). Multiple chicken (Gallus gallus) genome references to advance genetic variation studies.: Single haplotype chicken genome assemblies. *Cytogenetic and Genome Research*, *162*, 405–528. https://doi.org/10.1159/000529376

Wei, L., Liu, Y., Dubchak, I., Shon, J., and Park, J. (2002). Comparative genomics approaches to study organism similarities and differences. *Journal of Biomedical Informatics*, *35*(2), 142–150. https://doi.org/10.1016/S1532-0464(02)00506-3

Wilson, D. F. (2017). Oxidative phosphorylation: Regulation and role in cellular and tissue metabolism. *The Journal of Physiology*, *595*(23), 7023–7038. https://doi.org/10.1113/JP273839

Wu, C. (2007). Focal Adhesion. *Cell Adhesion and Migration*, *1*(1), 13–18.

Wu, L., Jiao, X., Zhang, D., Cheng, Y., Song, G., Qu, Y., and Lei, F. (2021). Comparative Genomics and Evolution of Avian Specialized Traits. *Current Genomics*, *22*(7), 496–511. https://doi.org/10.2174/1389202923666211227143952

Wu, Y. (2021). Molecular phyloecology suggests a trophic shift concurrent with the evolution of the first birds. *Communications Biology*, *4*(1), Article 1. https://doi.org/10.1038/s42003-021-02067-4

Wu, Y., Yan, Y., Zhao, Y., Gu, L., Wang, S., and Johnson, D. H. (2021). Genomic bases underlying the adaptive radiation of core landbirds. *BMC Ecology and Evolution*, *21*(1), 162. https://doi.org/10.1186/s12862-021-01888-5

Xie, W., Wucherpfennig, K., and Patel, D. J. (2023). A structural platform for B cell receptor signaling. *Cell Research*, *33*(2), 95–96. https://doi.org/10.1038/s41422-022-00724-9

Yang, J., Ueharu, H., and Mishina, Y. (2020). Energy metabolism: A newly emerging target of BMP signaling in bone homeostasis. *Bone*, *138*, 115467. https://doi.org/10.1016/j.bone.2020.115467

Yau, W. W., and Yen, P. M. (2020). Thermogenesis in Adipose Tissue Activated by Thyroid Hormone. *International Journal of Molecular Sciences*, *21*(8), 3020. https://doi.org/10.3390/ijms21083020

Zhang, G., Li, B., Li, C., Gilbert, M. T. P., Jarvis, E. D., Wang, J., and Avian Genome Consortium. (2014). Comparative genomic data of the Avian Phylogenomics Project. *GigaScience*, *3*(1), 26. https://doi.org/10.1186/2047-217X-3-26

Zhang, G., Li, C., Li, Q., Li, B., Larkin, D. M., Lee, C., Storz, J. F., Antunes, A., Greenwold, M. J., Meredith, R. W., Ödeen, A., Cui, J., Zhou, Q., Xu, L., Pan, H., Wang, Z., Jin, L., Zhang, P., Hu, H., … Froman, D. P. (2014). Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, *346*(6215), 1311–1320. https://doi.org/10.1126/science.1251385

Zoghby, B. A., Ryan, P. G., Little, R. M., Reid, T., and Hockey, P. A. (2015). Seasonal changes in movement and habitat use by Southern Ground-Hornbills in the South African lowveld. *Ostrich*, *86*(1–2), 87–95. https://doi.org/10.2989/00306525.2015.1030794

Chapter four – Comparative genomic analysis of the Southern Ground Hornbill and the Abyssinian Ground Hornbill – Genetic variants and selection analysis

4.1 Abstract

Genetic variation is a fundamental aspect of biological diversity between species. This variation enables species to withstand and adapt to various environmental conditions and challenges, thereby ensuring their survival. To gain insight into genomic variations associated with species-specific traits within the Southern Ground Hornbill (SGH), we performed a comparative analysis of the SGH proteome with that of the Abyssinian Ground Hornbill (AGH). The SGH genome encoded 5,466 species-specific orthogroups, approximately two times more than the AGH. Additionally, the SGH genome encompassed 104 species-specific orthogroups showing signs of gene duplication (five times more protein copies) compared to the AGH which only had nine. This provides a basis for gaining new genes and new functions in the SGH and AGH. Genetic variant analysis identified > 22 million variants between the genomes of the SGH and AGH. Post-filtering and merging of the variants showed that 162 protein-coding genes had multiple high-impact variants, including those coding for the MHC class II antigen proteins. Selection analysis of protein-coding genes with high-impact variants, however, revealed that 92.47% of the genes were under negative selection. This has important implications towards the survival and fitness of the SGH, as many of these genes code for proteins involved in reproduction. Ultimately the genetic diversity observed between the SGH and AGH may have provided these species with a selective advantage in adapting to their respective habitat south and north of the equator in the African savannah biome.

4.2 Introduction

Genetic variation is crucial for the persistence of species in a changing world with a plethora of challenges, including emerging diseases, climate change, geographical fragmentation and resource competition (Chung *et al*., 2023; Saccheri *et al*., 1998; Schoville *et al*., 2012). It allows for the introduction of biological diversity between species, which in turn allows them to evolve and adapt to changing conditions, thereby improving their chances of survival and reducing the risk of extinction (Chung *et al*., 2023; Höglund, 2009; Maia and Campos, 2021). Galapagos finches are a prime example of this phenomenon in avian species. They highlight the remarkable impact of genetic variation on the resilience of these species to adapt to different geographical regions and ecological conditions, through the diversification of their beak morphology and diet (Grant *et al*., 1976; Han *et al*., 2017).

Genetic variations can occur in the form of single nucleotide polymorphisms (SNPs), which involve single nucleotide changes within a genome and via large-scale genomic rearrangements in the form of structural variants (SVs) (Ruigrok *et al*., 2022). SNPs, a common form of genetic variation, can affect gene expression levels and protein functions through synonymous (no change in amino acid composition) and non-synonymous (change in a single amino acid within the translated protein) variations (Hunt *et al*., 2009; Ng and Henikoff, 2006). SNPs can thus cause phenotypic changes within species and influence their susceptibility to changes in environmental factors and stressors (Ng and Henikoff, 2006). While many non-synonymous mutations are deleterious and are often lost during purifying selection, some SNPs, particularly those that are neutral or beneficial, may be retained and contribute to the species' genetic diversity (Jackson *et al*., 2015; Ng and Henikoff, 2006).

SNPs have been used as a popular marker to study genetic diversity amongst several avian species including the chicken (Cendron *et al*., 2020; Strillacci *et al*., 2017), geese (Gao *et al*., 2023), turkey (Aslam *et al*., 2012) and ducks (Kraus *et al*., 2011; Zhu *et al*., 2016). However, with advancements in sequencing technologies making it possible to access genome-wide data, there is a growing interest in determining the significance of SVs in the evolution and adaptation of species (Delmore *et al*., 2023; Wold *et al*., 2023). SVs are genomic rearrangements ranging from 50 base pairs to over megabase pairs in sequence, making them difficult to detect using short-read sequencing technologies that only span across approximately 150 base pairs (Delmore *et al*., 2023; Ho *et al*., 2020). These variants are classified into different types including copy number variants - which include events of DNA gain and loss through deletions, duplications and insertions - as well as rearrangement events through translocations and inversions (Ho *et al*., 2020; Mahmoud *et al*., 2019). Compared to SNPs, SVs can result in substantial phenotypic changes, altering genome architecture, gene function, regulation and dosage, morphological and physiological traits among species (Alhakami *et al*., 2017; Liu *et al*., 2019; Mahmoud *et al*., 2019; Wu *et al*., 2021; Zhang *et al*., 2022).

Understanding genetic variation can provide insight into the myriad adaptations and evolutionary processes that species may have undergone throughout their existence (Erickson *et al*., 2004; Rubin *et al*., 2022). Extensive knowledge regarding genetic variation and its implication towards evolution and adaptation is available for avian species such as the chicken (Tan *et al*., 2024; Zhang *et al*., 2022), great tit (da Silva *et al*., 2018; Kim *et al*., 2018; Silva, 2020), Eurasian blackcaps and crows (Delmore *et al*., 2023; Weissensteiner *et al*., 2020). In

hornbill species, studies have mainly focused on understanding genetic variations from a population genetics perspective using microsatellite markers and variations within the mitochondrial DNA (Chamutpong *et al*., 2009; Dalton and Kotze, 2011; Jarulis *et al*., 2018; Sammler *et al*., 2013, 2012b, 2012a; Theron *et al*., 2013). To date, no studies have looked at genetic variations using SNPs and SVs at a genome-wide scale. Here, we took advantage of the whole genome sequence of the Southern Ground Hornbill (SGH) obtained in Chapter 2 and that of the publicly available Abyssinian Ground Hornbill (AGH), to identify genomic differences between these two ground hornbill species.

4.3 Methods and materials

4.3.1 Comparison of the SGH and AGH unique and gene duplicated proteins

The orthologous proteins conserved between the SGH and AGH were identified using OrthoFinder v 2.5.5 (Emms and Kelly, 2019) in **Chapter 3 section 3.3.2**. Shared (core) and species-specific proteins were extracted and their functions were characterised using the eggNOG functional annotations (**Chapter 3 section 3.3.1**; Cantalapiedra *et al*., 2021). The KO terms annotated in the eggNOG annotations were used to construct the KEGG pathways with the KEGG mapper v 5 tool Reconstruct (Kanehisa *et al*., 2022; Kanehisa and Sato, 2020). Similarly, core and species-specific orthogroups with evidence of gene duplication were extracted. The functional implications of orthogroups that contained a five-fold greater number of gene copies in either the SGH or AGH were further analysed using the COG annotations and KO terms from the eggNOG functional annotation outputs.

4.3.2 Analysis of genetic variations

4.3.2.1 Variant calling

The raw sequence data (both PacBio long and Illumina short reads datasets) for the AGH is publicly available on the GenomeArk database (www.genomeark.org/). Both datasets were utilized as reference sequences for the alignment of the SGH raw sequence data as obtained in Chapter 2 and subsequent variant calling (Figure 4.1).

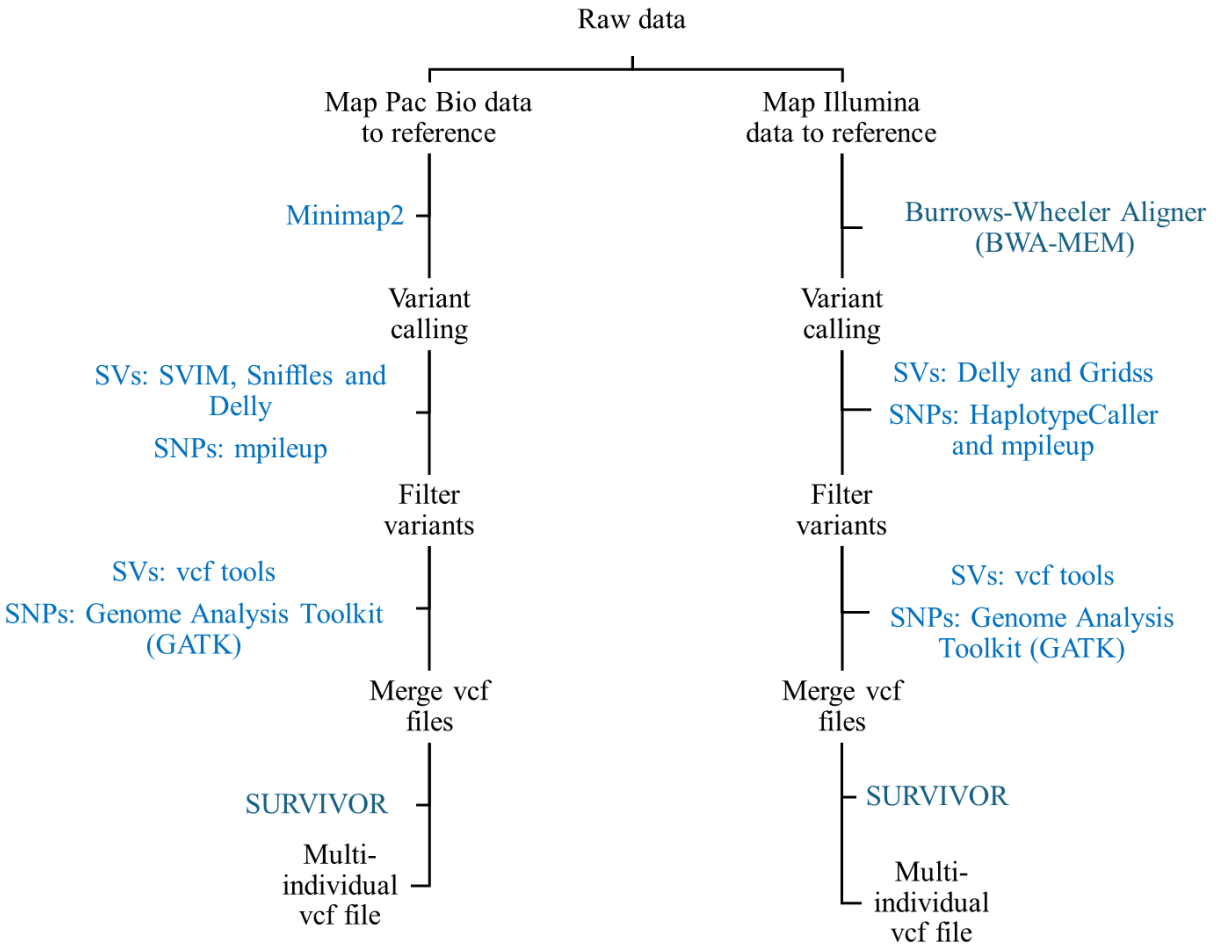


Figure 4.1: Workflow of the read-mapping based approach for variant calling within the SGH genome.

SVs can be detected using both read-mapping and *de novo* assembly-based strategies (Mahmoud *et al*., 2019). The read-mapping based strategy maps raw sequenced reads for SV calling, whereas the *de novo* assembly-based strategy uses the mapped raw reads to identify potential regions with SV signatures and performs *de novo* assembly of these regions before SV calling for a more comprehensive characterization of the SVs (Lin *et al*., 2023; Mahmoud *et al*., 2019). The calling step for both these strategies is similar and incorporates the identification of SV signatures from abnormally aligned reads in the form of intra- and inter-read alignments (Lin *et al*., 2023). Intra-read alignments are reads that span the entire SV position and result in the identification of signatures of insertions and deletions (Lin *et al*., 2023, 2022). Inter-read alignments are obtained from reads that align to larger SVs (resulting in supplementary alignments) and, based on differences in orientation, location and size of the reads, can be used to identify signatures of translocations, deletions, duplications and

inversions (Lin *et al*., 2023, 2022). Using the identified SV signatures, variant callers then cluster the SVs using different heuristic strategies, merge similar signatures from the abnormal alignments together and determine the likelihood of the SVs occurring (Lin *et al*., 2023). Similar to SVs, SNPs and small Insertion/Deletions (INDELs) can also be detected using read-mapping or *de novo* assembly-based strategies (Olson *et al*., 2015).

Short read variant calling

The quality-controlled and trimmed SGH short reads obtained from **Chapter 2 section 2.3.3** were mapped against the AGH short reads using the Burrows-Wheeler Aligner (BWA-MEM) v 0.7-17-r1188 (Li, 2013; Li and Durbin, 2009) with default parameters. Mapped reads were then converted into BAM format and sorted and indexed using SAMtools v 1.13 (Danecek *et al*., 2021). SV calling was conducted by a consensus-based approach with the parallel use of the SV calling tools Delly v 1.1.6 (Rausch *et al*., 2012) and Genomic Rearrangement Identification Software Suite (GRIDSS) v 2.13.2 (Cameron *et al*., 2017). This was done to achieve a reliable dataset of SVs with increased precision and maximize the types of SVs detected from the raw reads (Liu *et al*., 2022). Delly used the read-mapping based approach (Rausch *et al*., 2012), while GRIDSS used a *de novo* assembly-based approach with the assembler SPAdes (Bankevich *et al*., 2012; Cameron *et al*., 2017). Delly and GRIDSS characterized SVs in the form of deletions, duplications, inversions and translocation with GRIDSS further characterizing INDELs (Cameron *et al*., 2017; Rausch *et al*., 2012).

Since Delly and GRIDSS are only sensitive to large-genomic rearrangements, SNP calling was conducted using the HaplotypeCaller function from the Genome analysis tool kit (GATK) v 4-4.0.0 (McKenna *et al*., 2010) and the *mpileup* function from BCFtools v 1.13.1 (Li *et al*., 2009) with default parameters. These functions also identified the presence of small INDELs within the mapped reads. The HaplotypeCaller function used a *de novo* assembly-based approach through local *de novo* assembly of these candidate regions (Van der Auwera *et al*., 2013), while the *mpileup* function used a read-mapping based approach (Li *et al*., 2009).

Long read variant calling

The corrected and trimmed long reads generated using the first two steps of the Canu assembly in **Chapter 2 section 2.3.4** were mapped against the AGH long reads using Minimap2 (Li, 2018) with default parameters. Mapped reads were then converted into BAM format and sorted and indexed using SAMtools v 1.13 (Danecek *et al*., 2021). Similar to short read variant calling,

a consensus-based approach was used for SV calling with the tools Delly v 1.1.6 (Rausch *et al*., 2012) using the *delly_lr* option for long reads, Sniffles v 2.2 (Smolka *et al*., 2024) and SVIM v 2.0.0 (Heller and Vingron, 2019) using default parameters. The three SV callers used for the long reads all functioned using the read-mapping-based approach for SV calling. Sniffles and SVIM characterised SVs in the form of insertions, deletions, duplications, inversions and translocations (Heller and Vingron, 2019; Smolka *et al*., 2024).

Once again, since Delly, Sniffles and SVIM are sensitive to large-scale genomic rearrangements, SNP calling was conducted using the *mpileup* function from BCFtools using the *pacbio-ccs* option (Danecek *et al*., 2021). This allowed for the identification of small INDELs within the mapped reads.

4.3.2.2 Filtering and merging of variants

SV filtering

The SVs obtained from short and long-read variant calling were filtered to obtain a variant dataset comprised of high-confidence variants. SVs were filtered with a quality filter with the flag *PASS* (representing that all the parameters set by the variant caller for a high-confidence variant were met) and a minimum quality score of 30, using the *remove-filtered-all* and *minQ* commands from VCFtools v 0.1.16 (Danecek *et al*., 2011), respectively.

The filtered SVs determined using both the short- and long-read variant callers were merged with SURVIVOR v 1.0.3 (Jeffares *et al*., 2017). The options (1,000 1 0 0 0 50) used for SURVIVOR merging were 1) the SVs had to have a maximum distance of 1,000 bp between the breakpoints of the SVs (parameter value = 1,000), 2) the SVs had to be present in a minimum at least one variant caller (parameter value = 1), 3) the type of SV, strand they occurred on and the estimated distance between SVs based on their size were not taken into consideration (parameter values = 0) and 4) the SVs needed to be at least 50 bp in size (parameter value = 50). Merging of the filtered variants was performed independently for the short and long reads and ultimately, variants from both read datasets were merged to create a single vcf file for downstream analysis. To identify the total number of SVs, the ‘minimum number of variant callers’ criterion was also altered between one to five to determine the number of SVs commonly identified by all variant callers.

SNPs and small INDEL filtering

SNPs and small INDELs identified in the short- and long-read datasets were subjected to hard filtering using GATK (McKenna *et al*., 2010) with the recommended parameters 1) quality depth (QD) < 2, 2) quality (QUAL) < 30.0, 3) strand odds ratio (SOR) > 3.0, 4) Fisher strand bias (FS) > 60.0, 5) mapping quality (MQ) < 40.0, 6) MQ rank sum (MQRankSum) < -12.5 and 7) read position rank sum (ReadPosRankSum) < − 8.0. Hard filtering resulted in the SNPs and small INDELs receiving either a PASS flag if all the parameters were fulfilled or a flag with the name of the parameter that failed (Van der Auwera *et al*., 2013). The filtered SNPs and small INDELs were thus subjected to a final round of filtering using the remove-filtered-all command from BCFtools to retain only those with a PASS flag for merging. Since SURVIVOR is best suited for large SVs only (Jeffares *et al*., 2017), the SNPs and small INDELs datasets were independently merged to create a single vcf file for each variant type using the *concat* command from BCFtools (Danecek *et al*., 2021).

4.3.2.3 Annotation of variants

The filtered and merged variant datasets were sorted and indexed using BCFtools for annotation using SnpEff v 5.1f (Cingolani *et al*., 2012). There were over 20,000 pre-built databases of reference genomes available within SnpEff for annotations, however, no database of the AGH reference genome was available. Thus, a database using the genome sequence and annotations (from MOSGA, **Chapter 3 section 3.3.1**) of the AGH was built. Reference genome database annotations were then performed using the interval forest approach (Cingolani *et al*., 2012). In brief, each chromosome of the reference genome was assigned an interval tree composed of nodes representing the central point. Points of intervals on the right and left of the central point and points of intervals overlapping the central point were sorted according to the start and end positions (Cingolani *et al*., 2012). The input SGH vcf files were parsed through SnpEff and each variant was used as a query to identify intersecting genomic annotations. This was used to identify the genomic locations of the variants within the genome, their functional implications and their potential impacts on genes and proteins (Cingolani *et al*., 2012).

The SnpEff annotations were corroborated with the eggNOG annotations (COG categories and KEGG pathways) of the AGH (**Chapter 3 section 3.3.1**). This was done to identify the implications of variants between the SGH and AGH on the function of protein coding sequences.

4.3.3 Selection analysis

To identify the selective forces acting on protein-coding sequences in the presence of high-impact variants, the relative rate of nonsynonymous (dN) and synonymous (dS) nucleotide substitutions was assessed. This ratio is commonly used as a standard measure of selective pressures exerted on a gene over evolutionary time (Minias *et al*., 2018; Tine *et al*., 2016). A dN/dS ratio < 1 indicates negative (purifying) selection, in which deleterious alleles/traits are removed over time, and a dN/dS ratio > 1 indicates positive (diversifying) selection, in which advantageous alleles/traits are acquired (Nei and Gojobori, 1986; Yang and Nielsen, 2000).

The single-copy orthologous proteins conserved between the SGH and AGH identified by OrthoFinder v 2.5.5 (Emms and Kelly, 2019) in **Chapter 3 section 3.3.2** were used for selection analysis. In particular, those with the presence of high-impact variants were investigated. In brief, the sequences were aligned using ClustalW v 2.0 (Li, 2003). The dN/dS ratios were then calculated from the aligned sequences using codeml v 4.1 from the Phylogenetic analysis using the maximum likelihood (PAML) package (Álvarez-Carretero *et al*., 2023) implemented in JCoDA. Pairwise calculation of the dN/dS ratio was conducted using the sliding graph approach in JCODA v 1.4 (Steinway *et al*., 2010) with the Nei-Gojobori (1986) model (Nei and Gojobori, 1986).

4.4 Results

4.4.1 Species-specific differences between the SGH and AGH proteomes

The proteome of the SGH and AGH comprises a combined 67,520 proteins, which belong to 32,104 distinct orthogroups. The majority of these orthogroups (74.72%; 23,987 orthogroups) were shared between the two species. A further 5,466 orthogroups were specific to the SGH, almost twice as many as in the AGH (2,651 orthogroups). A total of 3,678 (11.46%) orthogroups show evidence of paralogy (proteins derived through gene duplication) with 85.07% of these orthogroups conserved in both birds. Of the core orthogroups, 1,679 and 552 orthogroups showed evidence of paralogy only in the AGH and SGH, respectively. Furthermore, 106 and 433 orthogroups incorporating paralogues were specific to the AGH and SGH, respectively.

4.4.1.1 Functional analysis of the core and species-specific orthogroups

Approximately 88.5% (comprising 51,161 proteins) of the core orthogroups were assigned to a functional COG supra-functional category ("information storage and processing", "cellular processes and signalling", "metabolism" and "poorly characterized functions") and twenty-seven COG categories. The core orthogroups of the SGH and AGH primarily incorporated proteins from the COG supra-functional category "poorly characterised function" (39.48%) (Figure 4.2). The core orthogroups were further dominated by proteins in the supra-functional categories "cellular processes and signalling" (29.31%) and "metabolism" (17.01%). Proteins in the supra-functional category "cellular processes and signalling" mainly functioned in signal transduction (14.64%) and post-translational modification, protein turnover and chaperone functions (6.01%) (Appendix Table 16). Within "metabolism", the proteins were mainly involved in amino acid metabolism and transport (7.89%), carbohydrate metabolism and transport (2.37%), inorganic ion transport and metabolism (1.94%) and lipid metabolism (1.64%).

A total of 71.33% (6,958 proteins) of the SGH- and AGH-specific orthogroups were assigned a functional COG category assignment. The species-specific orthogroups of the SGH (53.56%) and AGH (40.26%) primarily incorporated proteins belonging to the COG supra-functional category "poorly characterised function" (Figure 4.2). Additionally, the AGH-specific orthogroups had a greater abundance of proteins associated with the supra-functional categories "cellular processes and signalling" (26.56%), "information storage and processing" (17.55%) and "metabolism" (15.63%) compared to the SGH (18.06%, 14.60% and 13.78%, respectively). The difference in abundance is mainly due to the high number of proteins involved in "cellular processes and signalling" in the AGH which included signal transduction (12.23%), post-translational modification, protein turnover and chaperone functions (6.51%) and intracellular trafficking and secretion (3.00%) compared to the SGH (8.31%, 4.37% and 1.27%, respectively) (Appendix Table 16). Within "information storage and signalling", proteins mainly functioned in transcription (6.57%) and translation (1.87%) in the AGH compared to the SGH (4.58% and 0.53%, respectively).

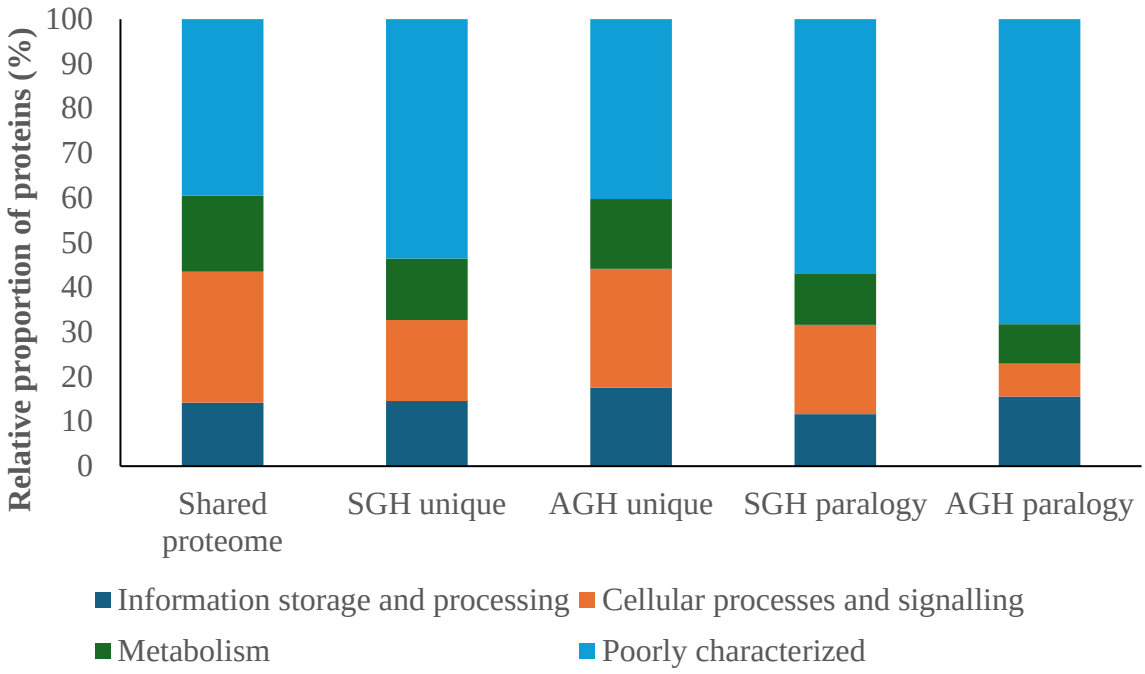


Figure 4.2: Relative proportion of core, species-specific and duplicated proteins in the SGH and AGH in each COG super-functional group. The SGH and AGH shared classification represents the orthogroups core to both species. The SGH and AGH paralogy classifications include orthogroups that were shared and unique to these species and that exhibited a five-fold difference in either of the taxon.

4.4.1.2 Functional analysis of the species-specific proteome using KEGG pathways

KO terms were assigned to 50.64% (25,909 proteins) and 24.69% (2,408 proteins) of the core and combined SGH- and AGH-specific orthogroups, respectively. The KO terms associated with the species-specific orthogroups were further used to reconstruct the KEGG pathways in which these proteins function. The AGH genome incorporates more species-specific proteins in the KEGG class "organism systems" (23.00%) compared to the SGH (20.84%). In contrast, the SGH genome incorporates more species-specific proteins functioning under the class "metabolism" (13.14%) compared to the AGH (12.57%) (Appendix Table 17). The KEGG pathways that exhibited a five-fold difference in the number of proteins identified in either the SGH- or AGH-specific orthogroups were further analysed to identify functional differences between the two species. This analysis mainly focused on the pathways involved in metabolism and immune response.

The species-specific proteome of the SGH and AGH exhibited variations within pathways under the categories of lipid, carbohydrate and nucleotide metabolism. In the lipid metabolism category, 9-fold and 7-fold more species-specific proteins involved in the glycerophospholipid and glycerolipid metabolism pathways, respectively are encoded on the SGH genome compared to the AGH (Appendix Table 18). These two pathways are required for energy production and the synthesis of structural and functional lipids utilised in the construction of cellular components (Chandel, 2021; Gyamfi *et al*., 2019). Within the carbohydrate metabolism category, the SGH genome coded for 5-fold more species-specific proteins involved in amino sugar and nucleotide sugar metabolism than the AGH genome (Appendix Table 18). This pathway plays a crucial role in synthesising amino and nucleotide sugars. Amino sugars (amino groups replace one or more hydroxyl groups in the sugar) form part of complex polysaccharides such as glycoproteins, glycolipids, and glycosaminoglycans, while nucleotide sugars (a nucleotide is attached to the anomeric carbon of the sugar) are intermediates for the biosynthesis of carbohydrates (Mikkola, 2020; Yang *et al*., 2023).

The pyrimidine metabolism pathway (part of nucleotide metabolism) synthesises pyrimidine nucleotides, which serve as the building blocks of nucleic acids. Additionally, they act as intermediates in energy transfer and metabolism, facilitating various biochemical reactions within the cell (Micheli *et al*., 2011). The SGH genome coded for seven species-specific proteins functioning in pyrimidine metabolism, which were absent in the AGH-specific orthogroups (Appendix Table 18). In avian species such as chickens, dietary nucleotide supplementation has been implicated in improving growth performance and immune response when under stressful conditions (Abdel Aziz *et al*., 2024; Jung and Batal, 2012; Salah *et al*., 2019). These differences in the unique proteome related to metabolism in the SGH and AGH could reflect variations in the physiology and adaptability of these species to their respective environments.

The AGH genome coded for 6-fold more species-specific proteins functioning under the complement and coagulation cascade pathway in the immune system than the SGH (Appendix Table 19). This pathway incorporates both the complement and coagulation cascades, which are proteolytic systems (capable of forming peptide bonds in proteins) in the blood (Amara *et al*., 2008). These two cascades play an essential role in innate immunity, the body's first line of defence and homoeostasis against pathogens and injurious stimuli (Amara *et al*., 2008). Although not directly studied in avian species, the dysregulation of this pathway has been

implicated in neurodegenerative diseases, inflammatory disorders and increased disease severity in some cases (Ma *et al*., 2021; Satyam *et al*., 2019; Zhang *et al*., 2022).

In the class "cellular processes", the efferocytosis pathway was also identified to mediate immune responses. Efferocytosis is the process of removing apoptotic (dying) cells via phagocytosis - the intake of extracellular substances by phagocytes (Moon *et al*., 2023). This process is required for tissue repair, resolving inflammation and homeostasis of the immune system (Ge *et al*., 2022). A total of eight species-specific proteins are encoded on the SGH genome functioning in efferocytosis (Appendix Table 19). The dysregulation of this pathway has been linked to cell necrosis and tissue damage via the accumulation of apoptotic cells and ultimately to various inflammatory disorders (Ge *et al*., 2022).

4.4.2 Gene duplications between the SGH and AGH

Orthogroups showing evidence of paralogy were analysed further. Particularly, those where greater than 5-fold protein copies were found in either the SGH or AGH relative to the other taxon were considered. This included 263 core (69.52% of these associated with the AGH genome), 104 SGH-specific (1,047 proteins) and nine AGH-specific orthogroups (65 proteins).

Approximately 86% (comprising 6,211 proteins) of the paralogous orthogroups (including the core and species-specific orthogroups with paralogues) were assigned a functional COG category. Analysis of the COG supra-functional classification of the paralogous orthogroups with 5-fold more protein copies showed that the majority of these proteins in the AGH belonged to the "poorly characterised function" (68.26%) compared to the SGH (56.84%) (Figure 4.2). By contrast, the SGH incorporated a greater proportion of orthogroups with 5-fold more protein copies associated with "cellular processes and signalling" (20.05%) and "metabolism" (11.51%), while only ~16% in total were associated with these COG supra-functional categories in the AGH (Appendix Table 16). The paralogous proteins over-represented in the SGH primarily functioned in signal transduction (8.31%), whereas it only contributed 1.57% in the AGH. Additionally, orthogroups with proteins functioning in post-translational modification, protein-turnover and chaperone functions (7.01%), energy production and conservation (4.41%), defence mechanisms (2.65%) and transcription (2.18%) were overrepresented in the SGH, while only 5.08%, 0.15%, 0.06% and 0.82% of the AGH proteins involved in these functional categories were overrepresented, respectively.

A total of 13.40% (832 proteins) of the paralogous orthogroups in the SGH and AGH were assigned a functional KO term. Construction of the KEGG pathways using these KO terms revealed that neither the SGH nor the AGH showed a 5-fold difference in the number of paralogous orthogroups associated with any of the reconstructed pathways when compared to each other. The threshold was thus reduced to a 3-fold difference in the number of paralogous orthogroups. This showed that two paralogous orthogroups incorporating three proteins functioning in "carbohydrate metabolism" were overrepresented in the SGH compared to the AGH, which had no paralogous orthogroups functioning in this category (Appendix Table 20). Additionally, the SGH had four paralogous orthogroups with proteins functioning in the regulation of the actin cytoskeleton pathway under "cell motility" compared to the AGH which only had one paralogous orthogroup with proteins functioning in this pathway. This pathway plays a role in various cellular processes including cell motility, adhesion, proliferation and cell shape maintenance via the formation of structures such as the lamellipodia, filopodia, focal adhesions and stress fibres (Carpenter, 2000; Lee and Dominguez, 2010).

The AGH had three paralogous orthogroups associated with the category glycan biosynthesis and metabolism, while the SGH only had one paralogous orthogroup functioning in this pathway (Appendix Table 20). The paralogous proteins identified in the AGH mainly functioned in the biosynthesis of N-glycans. N-glycans are glycans (carbohydrates composed of several sugar molecules) that are attached to the amide nitrogen of asparagine via N-glycosylation (Bieberich, 2014). N-glycans play an essential role in signal transduction, cell proliferation and regulating immune responses (Chao *et al*., 2020). An additional three paralogous orthogroups were also identified in the AGH functioning in the neutrophil extracellular trap formation pathway under the immune system compared to the SGH, which only had one paralogous orthogroup functioning in this pathway. Extracellular neutrophil traps, made up of chromatin coated with histones, proteases and neutrophil granule proteins, are stimulated by microorganisms and endogenous stimuli, which are subsequently immobilized and destroyed (Boeltz *et al*., 2019; Rada, 2019).

4.4.3 Analysis of genetic variations

4.4.3.1 Variant calling and filtering

A consensus-based approach involving multiple variants calling tools was performed to identify potential variants within the genomes of the SGH and AGH. Variant calling predicted 22,911,484 variants between the SGH and AGH genomes, of which 99.15% were accounted for by SNPs (including small INDELs) and only 0.85% by SVs (Table 4.1). A greater proportion of SVs and SNPs were identified within the short read dataset (73.20% and 61.34%) compared to the long read dataset (26.80% and 38.66%), respectively. The identified variants were subjected to filtering to achieve high-confidence variant datasets per caller. Approximately 34% and 89% of the identified SVs and SNPs were retained, respectively. The short-read dataset retained a greater proportion of SVs and SNPs (57.33% and 64.29%) compared to the long-read dataset (42.67% and 35.71%), respectively. The long-read variant callers Delly (90.11%) and Sniffles (88.04%) retained a greater proportion of the SVs compared to the short-read variant callers Delly (47.87%) and GRIDSS (24.55%). Long-read sequencing technologies have the advantage of spanning entire repetitive regions, which significantly improves the accuracy of alignment and variant calling (De Coster *et al*., 2021). In contrast, short reads often struggle in these regions due to their limited length, leading to fragmented detection of variants (Helal *et al*., 2024).

Table 4.1: Number of variants detected from the Illumina short reads and PacBio long reads using a consensus-based approach

	SVs			**SNPs and small INDELs**		
Read type	Variant callers	Identified	Retained	Variant callers	Identified	Retained
Short reads	Delly	12,602	6,032	HaplotypeCaller	6,934,656	6,696,260
	GRIDSS	129,906	31,894	mpileup	7,000,897	6,307,896
Long reads	Delly	16,848	15,182	mpileup	8,781,240	7,224,616
	SVIM	24,454	3,461			
	Sniffles	10,881	9,580			
Total		**194,691**	**66,149**		**22,716,793**	**20,228,772**

SVs obtained from the long reads comprised 14,241 deletions, 11,561 insertions, 129 duplications, 232 inversions and 2,060 break-ends (representing translocations). In contrast, SVs obtained from the short reads comprised 3,940 deletions, 649 insertions, 310 duplications, 647 inversions and 32,380 break-ends (Table 4.2). GRIDSS in particular was only able to characterise SVs into break-ends, which represented 85.38% of the SV types identified from the short read data. Due to this, the proportion of translocations was the highest among all the SV types identified. SNPs retained after filtering from the short and long reads were predominated by SNPs (88.60% and 78.77%) compared to small INDELs (11.40% and 21.23%), respectively.

The filtered SVs from each variant caller were merged using SURVIVOR (Jeffares *et al*., 2017). Despite having identified a greater proportion of SVs, only 2,606 SVs were retained for the short reads. A total of 45 SVs were found to be overlapped by both Delly and GRIDSS. In contrast, 11,149 SVs were retained for the long reads after merging. The long-read variant callers further had a greater number of overlaps, with 4,921 SVs being common to two callers and 922 being common to all three long-read variant callers used.

The short and long read variants were further merged and comprised 12,555 SVs (3,774 deletions, 4,559 insertions and 4,222 break-ends). Further, the merged short/long read SNP and small INDELs comprised 2,481,095 small INDELs and 6,939,970 SNPs, respectively.

Table 4.2: Different SV types detected from the Illumina short reads and PacBio long reads pre- and post-filtering

		Pre-filtering					**Post-filtering**				
Read type	SV caller	Break-end (Translocation)	Deletion	Duplication	Insertion	Inversion	Break-end (Translocation)	Deletion	Duplication	Insertion	Inversion
Short reads	Delly	2,334	5,891	1358	992	2,027	486	3,940	310	649	647
	GRIDSS	129,906	0	0	0	0	31,894	0	0	0	0
Long reads	Delly	2,578	7,872	119	6,072	207	1,650	7,314	106	5,923	189
	SVIM	3,612	9,310	179	11,268	85	118	1,994	11	1,337	1
	Sniffles	466	5,481	20	4,817	97	292	4,933	12	4,301	42
Total		**138,896**	**28,554**	**1,676**	**23,149**	**2,416**	**34,440**	**18,181**	**439**	**12,210**	**879**

4.4.3.2 Characterisation of variants by location and impact

The filtered and merged variants were annotated against the annotated AGH using SnpEff. Genes that comprise multiple transcripts can result in multiple annotations and effects being noted for the same variant causing an increase in the number of annotations predicted (Cingolani *et al*., 2012). As such, 46.44% more SVs were annotated than what was retained post-merging from the combined short- and long-read SV datasets (Table 4.3). In contrast, only 0.47% and 3.59% more SNPs and small INDELs were annotated than those retained from the merged short- and long-read SNP dataset. Quality analysis of the annotated SVs, SNPs and small INDELs revealed the presence of several warnings during the annotation process with small INDELs accounting for the most warnings. This may be indicative of potential annotation errors or inconsistencies identified between the input variants and the reference genome of the AGH (Cingolani *et al*., 2012). Despite the large number of warnings, zero errors were noted in the annotations of the SVs, SNPs and small INDELs.

Table 4.3: Annotation statistics of the genomic variants determined using SnpEFF

	SVs	**SNPs**	**Small INDELs**
Variants in the merged file	12,555	6,939,970	2,481,095
Variants annotated	23,439	6,973,062	2,573,570
Warnings received	1,287	7,557	133,258
Annotated variants retained post-filtering	23,439	6,964,719	2,570,345
Number of known variants (i.e. non-empty ID)	19,297	0	0
Number of multi-allelic VCF entries (i.e. more than two alleles)	0	33,092	92,475
Number of effects predicted	50,348	9,220,133	3,344,348
Variant rate	1 variant every 48,309 bases	1 variant every 162 bases	1 variant every 440 bases

SVs were found to be biallelic (only present in two alleles, the AGH used as a reference and the SGH represented as the query) with 82.33% being known variants (Table 4.3). The known variants are SVs that have been previously identified and are present in the SnpEff database built using the AGH annotations. In contrast, 0.48% of the SNPs and 32.90% of the small INDELs provided in the vcf file were found to be multi-allelic with none being known variants. Based on the genome size of the AGH (1.13 Gb), small INDELs had the greatest variant rate

per base pair compared to SNPs, and SVs had the lowest variant rate per base pair. Analysis of the distribution of variants across the various chromosomes (relative to the AGH reference genome) highlighted that the larger chromosomes (46.84 Mb to 136.45 Mb) tended to have a greater abundance of variants (> 4.45%) compared to the shorter chromosomes (2,790 kb to 4544.08 kb – < 4.10%). Chromosomes 1-3 (CM020024.1 - CM020027.1) and chromosome Z (CM020063.1), in particular, had the largest chromosome sizes (46.84 Mb to 136.45 Mb) and accounted for 36.99%, 34.66% and 35.62% of the total SVs, SNPs and small INDELs identified, respectively.

SnpEff further predicted the location of variants within each gene and calculated the coding effect of variants on protein-coding genes and the overall protein function (Cingolani *et al*., 2012). An average of 97.90% (12,350,427 variants) of the SVs, SNPs and small INDELs annotated were located within the non-coding regions of the gene, with 73.62% (9,286,772 variants) occurring within the intergenic and intron regions (Figure 4.3a). SVs, in particular, were predominant within the intergenic region, while SNPs and small INDELs were predominant within introns compared to SVs. In total, only 2.10% (264,402 variants) of the SVs, SNPs and small INDELs occurred within the exons of the genes, with SNPs (220,214 variants) occurring at a greater abundance compared to the small INDELs (42,676 variants) and SVs (1,512 variants).

SNPs and small INDELs mainly had a “modifier” impact on the protein-coding genes (Figure 4.3b). Variants with a modifier impact typically include non-coding variants or variants that influence the non-coding regions of a gene. These are challenging to predict and often lack clear evidence of their impact (Ensembl, 2024). Modifier SNPs (97.43%) and small INDELs (98.47%) were mainly within the intergenic, intron, upstream and downstream regions of the gene (Appendix Table 21). A total of 57.15% and 41.37% of the predicted SNPs were classified as missense mutations (causing a change in the amino acid sequence of the protein encoded by the gene) and silent mutations (causing no change in protein sequence), respectively. Additionally, 1.48% were classified as nonsense mutations that create a stop codon within the protein-coding sequence. SNPs introduced a premature stop codon within the protein gene reading frames in 0.04% of all cases (Appendix Table 21). This is considered a high-impact change as per the SnpEff and Ensembl impact rating (Cingolani *et al*., 2012; Ensembl, 2024).

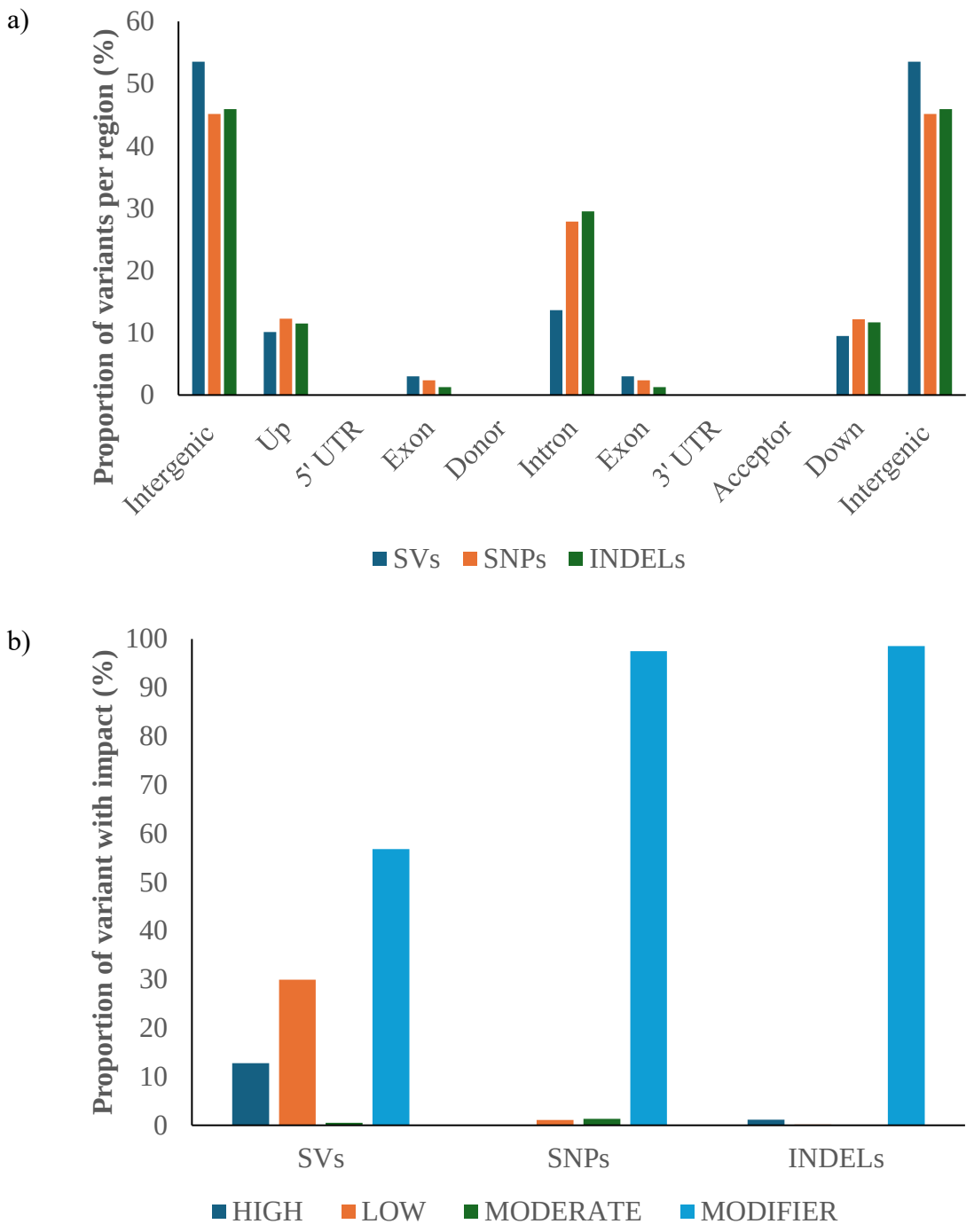


Figure 4.3: Genomic variants identified between the SGH and AGH are classified according to the a) region in the gene and b) the impact of the variant on the protein-coding sequence.

A large proportion of SVs had a modifier and low impact on protein-coding genes. However, 12.77% of SVs did have a high impact on the protein-coding region compared to SNPs and small INDELs, with only 0.62% average high-impact variants (Figure 4.3b). The types of impacts these SVs could have included transcript ablation, which is the deletion of a region containing a transcript, gene fusion and frameshift variants that lead to the disruption of the reading frame during protein translation due to the insertion or deletion of nucleotides (Cingolani *et al*., 2012; Ensembl, 2024) (Appendix Table 21). These can severely affect the function of the protein encoded by the variant gene copy.

4.4.3.3 Functional implications of high-impact variants

A total of 73,966 transcripts of protein-coding sequences were identified to have variants with an impact (low, moderate, modifier and high). Of these, 30.57% incorporated high-impact variants. Most of the transcripts with high-impact variants (87.43%; 19,769 proteins) were assigned a functional COG category. Analysis of the COG supra-functional categories highlighted that an average of 44.57% of the transcripts with high-impact variants had a "poorly characterized function", followed by 25.67% involved in "cellular processing and signalling" (Figure 4.4). The variants observed in genes involved in the latter category were mainly small INDELs (32.16%), followed by SVs (23.31%) and SNPs (21.52%). In contrast, SVs (22.08%) were predominant in the transcripts associated with the COG-supra-functional category "information storage and processing" (SNPs: 11.68% and small INDELs: 14.94%), while SNPs (16.22%) were predominant in transcripts associated with "metabolism" (small INDELs: 15.63% and SVs: 8.75%).

The transcripts with high-impact variants were mainly associated with the COG functions of signal transduction (12.57%), DNA replication, recombination and repair (8.09%), amino acid metabolism and transport (6.39%) and post-translational modification, protein turnover and chaperone functions (5.93%) (Appendix Table 22). Transcripts associated with the COG function of signal transduction had the highest abundance of small INDELs with a high impact (16.04%), while those associated with DNA replication and repair had the greatest abundance of SVs with a high impact (14.56%).

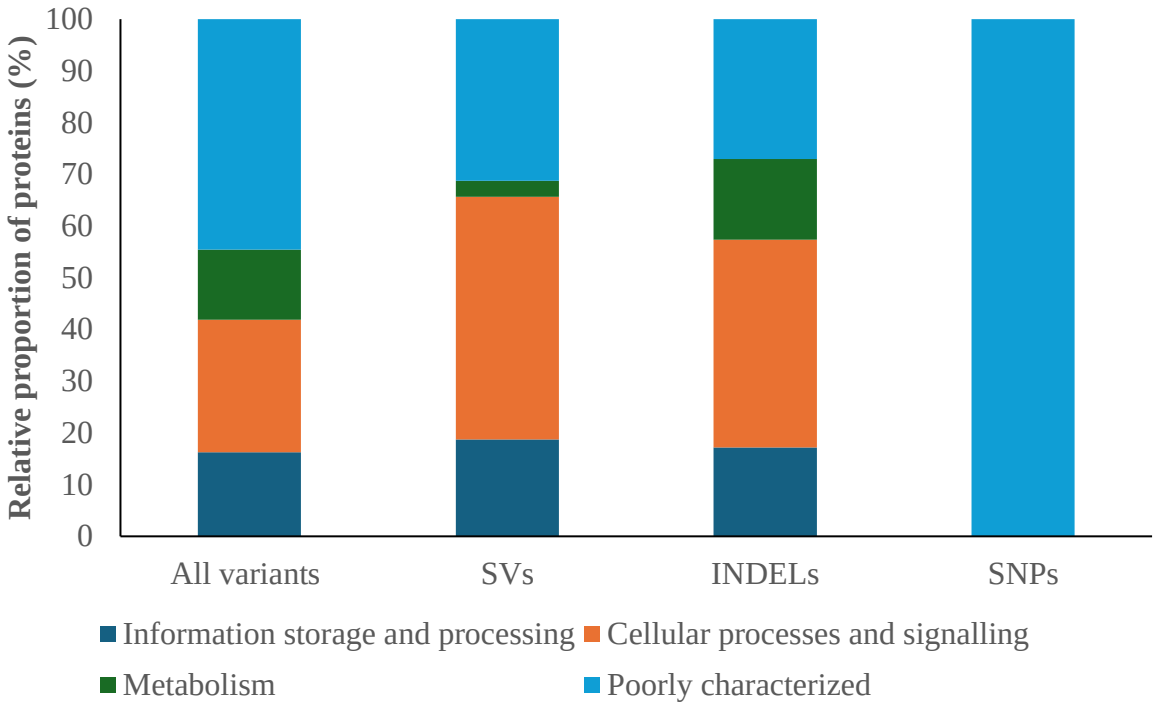


Figure 4.4: Relative proportion of proteins with genomic variants between the SGH and AGH in each COG super-functional group. The COG functional categories identified within all the genomic variations identified are indicated under all variants, this is followed by the COG functional categories of protein-coding sequences with ≥ 15 high-impact variants.

Since a single transcript of a protein-coding sequence can have multiple high-impact variant annotations associated with it, a minimum threshold was set and the functional implications of transcripts with ³ 15 high-impact variants were determined. In total, 162 protein-coding sequences (0.72%) met this criterion, of which 149 were assigned a functional COG category. The majority of these were associated with SVs and small INDELs, while transcripts with SNPs only accounted for 3.09% (5 protein-coding sequences). Transcripts with SVs and small INDELs were mainly associated with the COG supra-functional category "cellular processes and signalling" (43.53%) (Figure 4.4). Transcripts with ³ 15 high-impact SVs and small INDELs were mainly associated with the COG functions of post-translational modification, protein turnover and chaperone functions (25.00% and 7.38%) and signal transduction (15.63% and 14.75%), respectively.

KO terms could only be linked to 14/162 of the protein-coding sequences with high-impact variants, but as KO terms may be linked to multiple pathways, 30 KEGG pathways could be reconstructed. The majority of transcripts with ≥ 15 high-impact SVs were associated with

"human diseases" (56.67%) and "organism systems" (20%) (Appendix Table 23). Two of these transcripts code for major histocompatibility complex (MHC) class II molecules. These molecules play a crucial role in presenting exogenous peptide antigens to $CD4^+$ T cells to initiate the adaptive immune response (Arase, 2021). The adaptive immune response is a more specific immune response compared to the innate immune response and is developed over time as a response to pathogen exposure (Marshall *et al*., 2018; Minias *et al*., 2023).

A total of 117 KEGG pathways were reconstructed using 118 transcripts with ³ 15 high-impact small INDELs. These were primarily associated with "human diseases" (35.06%), "organism systems" (18.39%), "environmental information processing" (16.09%) and "metabolism" (12.64%). Transcripts coding for the proteins dishevelled segment polarity protein 3, cAMP response element binding protein (CREB)-binding protein and solute and solute carrier family 8 member A2 were identified to function in 55 of the reconstructed pathways, including fifteen pathways related to signal transduction (Appendix Table 23). Signal transduction pathways allow cells to perceive and transmit extracellular signals through a complex interplay of signalling molecules to regulate necessary biological activities (Chen and Wu, 2012).

4.4.4 Selection analysis

Selection analysis was performed on 93 single-copy orthologous genes with ≥ 15 high-impact variants. A dN/dS ratio < 1 (indicating negative selection) was observed for 92.47% of the transcripts, while only 7.53% of the high-impact variant transcripts had a dN/dS ratio > 1 (indicating positive selection) (Appendix Table 24). Variant genes that are under negative selection mainly code for eleven proteins involved in signal transduction (12.50%), followed by six proteins involved in RNA processing and modification (6.82%) and transcription (6.82%), respectively.

Of the proteins whose genes incorporate ≥ 15 high-impact variants under negative selection, 79.07% had an associated KO term, with 35 involved in a functional KEGG pathway. Apart from proteins involved in the class "human diseases" (33.59%), several were also involved in pathways associated with "organism systems" (21.09%), "metabolism" (12.50%) and "genetic information processing" (11.72%). Amongst these classes, a gene coding for the protein steroid 17-alpha-hydroxylase (CYP17), which is linked to steroid biosynthesis and ovarian steroidogenesis, was identified (Appendix Table 24). This pathway is responsible for the synthesis of steroid hormones from cholesterol in the gonads, brain, pineal glands and adrenals. Steroid hormones also include sex steroids such as androgens, progesterone and estrogen

(Schlinger, 2015). Additionally, the gene coding for the Fanconi anemia group M protein associated with the Fanconi anemia pathway under the class "genetic information processing" was also identified to undergo negative selection. The Fanconi anemia pathway plays a role in repairing DNA interstrand crosslinks, which are highly toxic lesions in the genome (Ceccaldi *et al*., 2016). DNA interstrand crosslinks inhibit transcription and replication by preventing the DNA strand from separating (Deans and West, 2011).

A total of seven genes with ≥ 15 high-impact variants were found to be under positive selection. While no function could be assigned to most of these genes, one codes for a DC-STAMP domain-containing protein 2 which has been linked to osteoclast differentiation (**Chapter 3 section 3.4.4**) and immune regulation via inflammatory responses (Garcia-Hernandez *et al*., 2022). Another gene under positive selection codes for Retinal dehydrogenase 1, an oxidoreductase that catalyses the conversion of retinal to retinoic acid which has been identified to play a role in cell growth, differentiation and embryogenesis (Cho *et al*., 2021).

4.5 Discussion

4.5.1 The SGH and AGH evolved different repertoires of proteins related to metabolism and immune response

Comparative analysis of the AGH and SGH genomes revealed extensive differences at both the nucleotide and protein levels. While more than three-quarters of the proteins encoded on each genome are conserved in both species, 25.28% of the orthogroups are unique to either the AGH or SGH. The SGH genome in particular coded for 67.26% of all species-specific proteins. These findings suggest that after the divergence of the SGH and AGH from the Asian hornbills (Viseshakul *et al*., 2011), the genomes of the two *Bucorvus* species have undergone distinct gene gain and loss events, resulting in evolutionary changes at the protein level. Species-specific proteins can evolve via two important mechanisms, gene duplication followed by divergence and *de novo* gene birth from nongenic regions of the genome (Montañés *et al*., 2023). These changes may have helped them to adapt and cope with the new environmental challenges they may have encountered in their respective habitats.

The SGH genome exhibits a unique set of species-specific proteins involved in carbohydrate, lipid, and nucleotide metabolism, which were absent in the AGH. Differences in metabolism and metabolic rates could be associated with changes in environmental conditions, phenotypic

flexibility, life history traits and differences in diet between species (Auer *et al*., 2015; Barzegar *et al*., 2020; Jimenez *et al*., 2014; Norin and Metcalfe, 2019; Podlesak and McWilliams, 2006; White *et al*., 2007). Between the *Bucorvus* species, the major difference in physical conditions lies in its distribution across Africa. The SGH is found south of the equator, where they occur in large, vegetated grassland and savannah biomes, while the AGH is found north of the equator, in arid (dry and less rainfall) habitats (Kemp and Kemp, 1980; Mitchell *et al*., 1996; van Perlo, 1999). These different environmental conditions may require varying rates of energy consumption, thermogenesis and activity (Swanson *et al*., 2023). Furthermore, despite sharing a similar diet composed of insects, small mammals and reptiles (Mabula Ground-Hornbill Project, 2024), the availability and composition of food could vary between the two species within their respective environments.

Differences were also observed between the AGH and SGH in the protein datasets associated with immune responses. Variation in the immune response of avian species has been linked to several factors, including changes in environmental conditions, pathogen pressure and life-history trade-offs (Chang van Oordt *et al*., 2022; Hegemann *et al*., 2012; Horrocks *et al*., 2011; Minias *et al*., 2023). The AGH genome, in particular, coded for 6-fold more species-specific proteins functioning in the complement and coagulation cascade, which is associated with the innate immune response (Amara *et al*., 2008). These differences indicate that the two species might be experiencing distinct stresses in their respective environments. The AGH could have evolved to incorporate a robust first line of defence against different pathogens, while the SGH might have developed efficient cellular processes that regulate immune responses and prevent tissue damage (Ge *et al*., 2022). This divergence in immune response strategies between the two species could be a result of different environmental pressures, leading to unique evolutionary adaptations.

Almost double the number of orthogroups encompassing paralogous proteins are encoded on the AGH genome than that of the SGH. This suggests a more frequent occurrence of gene duplication events in the former species. Gene duplication is an important mechanism that gives rise to new genes that can either retain the original function or acquire new functions over time. This process creates genetic diversity between species, allowing them to adapt and evolve (Álvarez-Lugo and Becerra, 2021). While this phenomenon has been previously noted in the MHC class I and II of several avian species (Balakrishnan *et al*., 2010; Burri *et al*., 2010; Miller and Lambert, 2004; Minias *et al*., 2021), in other immune-related proteins functional

diversity is linked to gene conversion, rather than gene duplications (Lundqvist *et al*., 2006). In this case, gene duplications can also potentially create functional redundancy within the SGH and AGH. This redundancy could ensure that the proteins associated with essential processes remain flexible and resilient even in the event of mutations and environmental stresses, thereby aiding in their survival and adaptation (Crow and Wagner, 2006).

It should be noted that functional annotations of proteins within a genome provide insights into the presence and absence of these proteins between the species; however, it does not indicate whether these proteins are expressed or not (Brown, 2002). The actual function of these proteins may thus vary based on gene expression levels, post-translational modifications, and protein-protein interactions (Alberts *et al*., 2002; Lee and Yaffe, 2016). For example, analysis of the metabolic profiles of the black-capped chickadees (*Poecile atricapillus*) and American goldfinches (*Spinus tristis*) displayed variations in gene expression profiles related to lipid oxidation and amino acids metabolism in response to changes in diet and environmental conditions (Wone and Swanson, 2022). Thus, the integration of omics techniques, gene expression, and protein-protein interaction analysis of key proteins involved in adaption and survival mechanisms is crucial to identify the exact function of these proteins in the SGH and AGH.

4.5.2 SVs in immune-related proteins could affect the fitness and survival of the SGH

Over 22 million genomic variants were identified between the SGH and AGH, of which 41.17% were retained post-filtering and merging. Annotation of these variants highlighted that 97.78% of the SVs, SNPs and small INDELs identified were located within non-coding regions of the genes. Similar findings of genetic variants dominating intergenic and intron regions (> 98.00%) were observed in two chicken breeds (Fan *et al*., 2013).

Given the extensive size of the variant dataset achieved for the SGH and AGH, we mainly investigated the functional implications of genetic variation on those genes with ≥ 15 high-impact (13.39% of the total annotated variants) variants. The encoded proteins primarily play a role in post-translational modification, protein turnover and chaperone functions and signal transduction. Furthermore, two genes coding for MHC class II antigen proteins functioning in the adaptive immune response were also identified. High levels of genetic variations have been observed in the MHC class II antigens of several avian orders (Balasubramaniam *et al*., 2016;

He *et al*., 2020). In the SGH and AGH, this variation could positively or negatively influence their immune response towards foreign (e.g. from extracellular pathogens) and self-derived antigens (Hess and Edwards, 2002). It could either enhance the ability of the SGH or AGH to respond to pathogens (conferring resistance) or it could impair its function (resulting in a weak immune response) (Hess and Edwards, 2002). This could have significant implications for the survival and fitness of the two species, thus elucidating the exact impact of this genetic variation is essential via population genetic studies and comparative analysis with other hornbill species.

The impact of genetic variants on the species can vary depending on various factors such as its location within the gene, the type of variants and the specific protein sequence involved (Vihinen, 2015). For example, it has been observed that genetic variants within non-coding regions influence phenotypes such as beak morphology and plumage coloration in several avian species (Funk and Taylor, 2019; Wu *et al*., 2021; Yusuf *et al*., 2020). Furthermore, genetic variation disrupting regulatory regions by altering the spacing between cis-regulatory elements have also been linked to phenotypic changes in avian species (Abe and Gemmell, 2016; Strillacci *et al*., 2017). Thus, while this study has focused on high-impact variants, which are predicted to have a substantial impact on protein function, it should be interpreted with caution. Identifying which variant, in particular, is causing a phenotypic change irrespective of a low or high-impact classification is difficult to predict in a complex biological system comprised of a multitude of genes and proteins with interlinked functions (Vihinen, 2015). However, future studies should aim to incorporate all of these elements to provide a more comprehensive view of the genetic variation characterising potential phenotypic changes.

4.5.3 Proteins crucial for survival and reproduction are conserved via negative selection

Selection analysis of genes coding for single-copy orthologous proteins showed a strong prevalence of proteins under negative selection (~ 92%) compared to positive selection (~ 7.50%), indicating that the maintenance of existing gene function between the SGH and AGH is prioritized (Minias *et al*., 2018). In contrast, the few proteins under positive selection could be evolving to optimize their function and assist the SGH and AGH in adapting to their respective environments. Among the proteins undergoing negative selection, CYP17, belonging to the ovarian steroidogenesis and steroid biosynthesis pathways, is responsible for synthesizing the steroid hormone androgen from progestin and pregnane precursors via 17α-

hydroxy intermediates (Schlinger *et al*., 1999). It has been linked to the regulation of reproduction, sex determination, ovarian formation and follicular development in avian species (Freking *et al*., 2000; Huang *et al*., 2018). Furthermore, the Fanconi anemia complementation Group M protein identified within the Fanconi anemia pathway has also been associated with spermatogenic failure and premature ovarian failure in mice models (Fu *et al*., 2016; Yin *et al*., 2019). Avian species have been identified to exhibit similar sensitivity to what has been observed in mammals in the presence of deficient Fanconi anemia genes, thus these findings suggest that the conservation of these proteins and processes could be crucial for the survival and reproduction of the SGH and AGH (Liu *et al*., 2003; Nanda *et al*., 2007).

It should be noted that structural variant and selection analysis in this study is based on two genomes only and thus lacks comprehensive representation across populations of the two species. This study does, however, provide a glimpse into the genetic diversity of the SGH and AGH and creates a foundation for population-wide genomic analysis across different geographical locations. This will aid in enhancing our understanding of the adaptation, evolution and conservation potential of these two species.

4.6 Conclusion

This study provides a first glance into the genetic variation characterising the genomes of the SGH and AGH. Our analysis suggests that, despite belonging to the same genus, these two species have gained or lost a significant proportion of genes coding for proteins, conferring a selective advantage for them to adapt to their disparate environments, north (AGH) and south (SGH) of the equator. Additionally, genetic variation observed within protein-coding sequences and selection analysis of these genes underline the delicate balance between conserving essential biological functions and optimizing others in response to environmental pressures. It is important to corroborate these findings through population genetic analysis between multiple individuals of the AGH and SGH from different geographical locations across Africa to identify the exact impact of this variation on these species. Ultimately, the divergence of the SGH and AGH from Asian hornbills and their subsequent adaptation to their respective habitats in the African savannah biomes on opposite sides of the equator demonstrates the complex interplay between genetics, environment and evolution of these birds. As Ernst Mayr indicated: “There is probably no more original, more complex, and bolder concept in the history of ideas than Darwin's mechanistic explanation of adaptation.” (Mayr, 1982).

4.7 References

Abdel Aziz, A. A. M., Abdel Aziz, E. S. A., Khairy, M. H., Fadel, C., Giorgi, M., and Abdelaziz, A. S. (2024). The effect of butyric acid and nucleotides supplementation on broiler (*Gallus gallus domesticus*) growth performance, immune status, intestinal histology, and serum parameters. *Open Veterinary Journal*, *14*(1), 324–334. https://doi.org/10.5455/OVJ.2024.v14.i1.29

Abe, H., and Gemmell, N. J. (2016). Evolutionary Footprints of Short Tandem Repeats in Avian Promoters. *Scientific Reports*, *6*(1), 19421. https://doi.org/10.1038/srep19421

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). Analyzing Protein Structure and Function. In *Molecular Biology of the Cell. 4th edition*. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26820/

Alhakami, H., Mirebrahim, H., and Lonardi, S. (2017). A comparative evaluation of genome assembly reconciliation tools. *Genome Biology*, *18*, 93. https://doi.org/10.1186/s13059-017-1213-3

Álvarez-Carretero, S., Kapli, P., and Yang, Z. (2023). Beginner's Guide on the Use of PAML to Detect Positive Selection. *Molecular Biology and Evolution*, *40*(4), msad041. https://doi.org/10.1093/molbev/msad041

Álvarez-Lugo, A., and Becerra, A. (2021). The Role of Gene Duplication in the Divergence of Enzyme Function: A Comparative Approach. *Frontiers in Genetics*, *12*, 641817. https://doi.org/10.3389/fgene.2021.641817

Amara, U., Rittirsch, D., Flierl, M., Bruckner, U., Klos, A., Gebhard, F., Lambris, J. D., and Huber-Lang, M. (2008). Interaction Between the Coagulation and Complement System. *Advances in Experimental Medicine and Biology*, *632*, 71–79.

Arase, H. (2021). The major histocompatibility complex: New insights from old molecules into the pathogenesis of autoimmunity. *International Immunology*, *33*(12), 641–645. https://doi.org/10.1093/intimm/dxab079

Aslam, M. L., Bastiaansen, J. W., Elferink, M. G., Megens, H. J., Crooijmans, R. P., Blomberg, L. A., Fleischer, R. C., Van Tassell, C. P., Sonstegard, T. S., Schroeder, S. G., Groenen,

M. A., and Long, J. A. (2012). Whole genome SNP discovery and analysis of genetic diversity in Turkey (Meleagris gallopavo). *BMC Genomics*, *13*(1), 391. https://doi.org/10.1186/1471-2164-13-391

Auer, S. K., Salin, K., Rudolf, A. M., Anderson, G. J., and Metcalfe, N. B. (2015). Flexibility in metabolic rate confers a growth advantage under changing food availability. *Journal of Animal Ecology*, *84*(5), 1405–1411.

Balakrishnan, C. N., Ekblom, R., Völker, M., Westerdahl, H., Godinez, R., Kotkiewicz, H., Burt, D. W., Graves, T., Griffin, D. K., Warren, W. C., and Edwards, S. V. (2010). Gene duplication and fragmentation in the zebra finch major histocompatibility complex. *BMC Biology*, *8*(1), 29. https://doi.org/10.1186/1741-7007-8-29

Balasubramaniam, S., Bray, R. D., Mulder, R. A., Sunnucks, P., Pavlova, A., and Melville, J. (2016). New data from basal Australian songbird lineages show that complex structure of MHC class II β genes has early evolutionary origins within passerines. *BMC Evolutionary Biology*, *16*(1), 112. https://doi.org/10.1186/s12862-016-0681-5

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Barzegar, S., Wu, S. B., Choct, M., and Swick, R. A. (2020). Factors affecting energy metabolism and evaluating net energy of poultry feed. *Poultry Science*, *99*(1), 487–498. https://doi.org/10.3382/ps/pez554

Bieberich, E. (2014). Synthesis, processing, and function of N-glycans in N-glycoproteins. *Advances in Neurobiology*, *9*, 47–70. https://doi.org/10.1007/978-1-4939-1154-7_3

Boeltz, S., Amini, P., Anders, H. J., Andrade, F., Bilyy, R., Chatfield, S., Cichon, I., Clancy, D. M., Desai, J., Dumych, T., Dwivedi, N., Gordon, R. A., Hahn, J., Hidalgo, A., Hoffmann, M. H., Kaplan, M. J., Knight, J. S., Kolaczkowska, E., Kubes, P., … Herrmann, M. (2019). To NET or not to NET:current opinions and state of the science regarding the

formation of neutrophil extracellular traps. *Cell Death and Differentiation*, *26*(3), 395–408. https://doi.org/10.1038/s41418-018-0261-x

Brown, T. A. (2002). Understanding a Genome Sequence. In *Genomes. 2nd edition*. Wiley-Liss. https://www.ncbi.nlm.nih.gov/books/NBK21136/

Burri, R., Salamin, N., Studer, R. A., Roulin, A., and Fumagalli, L. (2010). Adaptive Divergence of Ancient Gene Duplicates in the Avian MHC Class II β. *Molecular Biology and Evolution*, *27*(10), 2360–2374. https://doi.org/10.1093/molbev/msq120

Cameron, D. L., Schröder, J., Penington, J. S., Do, H., Molania, R., Dobrovic, A., Speed, T. P., and Papenfuss, A. T. (2017). GRIDSS: Sensitive and specific genomic rearrangement detection using positional de Bruijn graph assembly. *Genome Research*, *27*(12), 2050–2060. https://doi.org/10.1101/gr.222109.117

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), 5825–5829. https://doi.org/10.1093/molbev/msab293

Carpenter, C. L. (2000). Actin cytoskeleton and cell signaling. *Critical Care Medicine*, *28*(4), N94.

Ceccaldi, R., Sarangi, P., and D'Andrea, A. D. (2016). The Fanconi anaemia pathway: New players and new functions. *Nature Reviews Molecular Cell Biology*, *17*(6), 337–349. https://doi.org/10.1038/nrm.2016.48

Cendron, F., Perini, F., Mastrangelo, S., Tolone, M., Criscione, A., Bordonaro, S., Iaffaldano, N., Castellini, C., Marzoni, M., Buccioni, A., Soglia, D., Schiavone, A., Cerolini, S., Lasagna, E., and Cassandro, M. (2020). Genome-Wide SNP Analysis Reveals the Population Structure and the Conservation Status of 23 Italian Chicken Breeds. *Animals: An Open Access Journal from MDPI*, *10*(8), 1441. https://doi.org/10.3390/ani10081441

Chamutpong, S., Saito, D. S., Viseshakul, N., Nishiumi, I., Poonswad, P., and Ponglikitmongkol, M. (2009). Isolation and characterization of microsatellite markers from the great hornbill, *Buceros* bicornis. *Molecular Ecology Resources*, *9*(2), 591–593. https://doi.org/10.1111/j.1755-0998.2008.02447.x

Chandel, N. S. (2021). Lipid Metabolism. *Cold Spring Harbor Perspectives in Biology*, *13*(9), a040576. https://doi.org/10.1101/cshperspect.a040576

Chang van Oordt, D. A., Taff, C. C., Ryan, T. A., and Vitousek, M. N. (2022). Timing of Breeding Reveals a Trade-Off between Immune Investment and Life History in Tree Swallows. *Integrative and Comparative Biology*, *62*(6), 1629–1639. https://doi.org/10.1093/icb/icac033

Chao, Q., Ding, Y., Chen, Z. H., Xiang, M. H., Wang, N., and Gao, X. D. (2020). Recent Progress in Chemo-Enzymatic Methods for the Synthesis of N-Glycans. *Frontiers in Chemistry*, *8*. https://doi.org/10.3389/fchem.2020.00513

Chen, B. S., and Wu, C. C. (2012). On the calculation of signal transduction ability of signaling transduction pathways in intracellular communication: Systematic approach. *Bioinformatics*, *28*(12), 1604–1611. https://doi.org/10.1093/bioinformatics/bts159

Cho, K., Lee, S. M., Heo, J., Kwon, Y. M., Chung, D., Yu, W. J., Bae, S. S., Choi, G., Lee, D. S., and Kim, Y. (2021). Retinaldehyde Dehydrogenase Inhibition-Related Adverse Outcome Pathway: Potential Risk of Retinoic Acid Synthesis Inhibition during Embryogenesis. *Toxins*, *13*(11), Article 11. https://doi.org/10.3390/toxins13110739

Chung, M. Y., Merilä, J., Li, J., Mao, K., López-Pujol, J., Tsumura, Y., and Chung, M. G. (2023). Neutral and adaptive genetic diversity in plants: An overview. *Frontiers in Ecology and Evolution*, *11*. https://doi.org/10.3389/fevo.2023.1116814

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., and Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Crow, K. D., and Wagner, G. P. (2006). What Is the Role of Genome Duplication in the Evolution of Complexity and Diversity? *Molecular Biology and Evolution*, *23*(5), 887–892. https://doi.org/10.1093/molbev/msj083

da Silva, V. H., Laine, V. N., Bosse, M., Oers, K. van, Dibbits, B., Visser, M. E., M. A. Crooijmans, R. P., and Groenen, M. A. M. (2018). CNVs are associated with genomic architecture in a songbird. *BMC Genomics*, *19*(1), 195. https://doi.org/10.1186/s12864-018-4577-1

Dalton, D., and Kotze, A. (2011). *Isolation and characterization of microsatellite loci in Southern Ground Hornbills (Bucorvus leadbeateri)*.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

De Coster, W., Weissensteiner, M. H., and Sedlazeck, F. J. (2021). Towards population-scale long-read sequencing. *Nature Reviews Genetics*, *22*(9), 572–587. https://doi.org/10.1038/s41576-021-00367-3

Deans, A. J., and West, S. C. (2011). DNA interstrand crosslink repair and cancer. *Nature Reviews Cancer*, *11*(7), 467–480. https://doi.org/10.1038/nrc3088

Delmore, K. E., Van Doren, B. M., Ullrich, K., Curk, T., van der Jeugd, H. P., and Liedvogel, M. (2023). Structural genomic variation and migratory behavior in a wild songbird. *Evolution Letters*, *7*(6), 401–412. https://doi.org/10.1093/evlett/qrad040

Emms, D. M., and Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. https://doi.org/10.1186/s13059-019-1832-y

Ensembl. (2024). *Ensembl Variation—Calculated variant consequences*. Ensembl. https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

Erickson, D. L., Hamrick, J. L., and Kochert, G. D. (2004). Ecological determinants of genetic diversity in an expanding population of the shrub Myrica cerifera. *Molecular Ecology*, *13*(6), 1655–1664. https://doi.org/10.1111/j.1365-294X.2004.02139.x

Fan, W. L., Ng, C. S., Chen, C. F., Lu, M. Y. J., Chen, Y. H., Liu, C. J., Wu, S. M., Chen, C. K., Chen, J. J., Mao, C. T., Lai, Y. T., Lo, W. S., Chang, W. H., and Li, W. H. (2013).

Genome-Wide Patterns of Genetic Variation in Two Domestic Chickens. *Genome Biology and Evolution*, *5*(7), 1376–1392. https://doi.org/10.1093/gbe/evt097

Freking, F., Nazairians, T., and Schlinger, B. A. (2000). The Expression of the Sex Steroid-Synthesizing Enzymes CYP11A1, 3β-HSD, CYP17, and CYP19 in Gonads and Adrenals of Adult and Developing Zebra Finches. *General and Comparative Endocrinology*, *119*(2), 140–151. https://doi.org/10.1006/gcen.2000.7503

Fu, C., Begum, K., and Overbeek, P. A. (2016). Primary Ovarian Insufficiency Induced by Fanconi Anemia E Mutation in a Mouse Model. *PLoS ONE*, *11*(3), e0144285. https://doi.org/10.1371/journal.pone.0144285

Funk, E. R., and Taylor, S. A. (2019). High-throughput sequencing is revealing genetic associations with avian plumage color. *The Auk*, *136*(4), ukz048. https://doi.org/10.1093/auk/ukz048

Gao, G., Zhang, H., Ni, J., Zhao, X., Zhang, K., Wang, J., Kong, X., and Wang, Q. (2023). Insights into genetic diversity and phenotypic variations in domestic geese through comprehensive population and pan-genome analysis. *Journal of Animal Science and Biotechnology*, *14*(1), 150. https://doi.org/10.1186/s40104-023-00944-y

Garcia-Hernandez, M. de la L., Rangel-Moreno, J., Garcia-Castaneda, M., Kenney, H. M., Paine, A., Thullen, M., Anandarajah, A. P., Schwarz, E. M., Dirksen, R. T., and Ritchlin, C. T. (2022). Dendritic cell-specific transmembrane protein is required for synovitis and bone resorption in inflammatory arthritis. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.1026574

Ge, Y., Huang, M., and Yao, Y. (2022). Efferocytosis and Its Role in Inflammatory Disorders. *Frontiers in Cell and Developmental Biology*, *10*, 839248. https://doi.org/10.3389/fcell.2022.839248

Grant, P. R., Grant, B. R., Smith, J. N., Abbott, I. J., and Abbott, L. K. (1976). Darwin's finches: Population variation and natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(1), 257–261. https://doi.org/10.1073/pnas.73.1.257

Gyamfi, D., Ofori Awuah, E., and Owusu, S. (2019). Chapter 2 - Lipid Metabolism: An Overview. In V. B. Patel (Ed.), *The Molecular Nutrition of Fats* (pp. 17–32). Academic Press. https://doi.org/10.1016/B978-0-12-811297-7.00002-0

Han, F., Lamichhaney, S., Grant, B. R., Grant, P. R., Andersson, L., and Webster, M. T. (2017). Gene flow, ancient polymorphism, and ecological adaptation shape the genomic landscape of divergence among Darwin's finches. *Genome Research*, *27*(6), 1004–1015. https://doi.org/10.1101/gr.212522.116

He, K., Minias, P., and Dunn, P. O. (2020). Long-Read Genome Assemblies Reveal Extraordinary Variation in the Number and Structure of MHC Loci in Birds. *Genome Biology and Evolution*, *13*(2), evaa270. https://doi.org/10.1093/gbe/evaa270

Hegemann, A., Matson, K. D., Both, C., and Tieleman, B. I. (2012). Immune function in a free-living bird varies over the annual cycle, but seasonal patterns differ between years. *Oecologia*, *170*(3), 605–618. https://doi.org/10.1007/s00442-012-2339-3

Helal, A. A., Saad, B. T., Saad, M. T., Mosaad, G. S., and Aboshanab, K. M. (2024). Benchmarking long-read aligners and SV callers for structural variation detection in Oxford nanopore sequencing data. *Scientific Reports*, *14*(1), 6160. https://doi.org/10.1038/s41598-024-56604-2

Heller, D., and Vingron, M. (2019). SVIM: Structural variant identification using mapped long reads. *Bioinformatics*, *35*(17), 2907–2915. https://doi.org/10.1093/bioinformatics/btz041

Hess, C. M., and Edwards, S. V. (2002). The Evolution of the Major Histocompatibility Complex in Birds: Scaling up and taking a genomic approach to the major histo compatibilty complex (MHC) of birds reveals surprising departures from generalities found in mammals in both large-scale structure and the mechanisms shaping the evolution of the MHC. *BioScience*, *52*(5), 423–431. https://doi.org/10.1641/0006-3568(2002)052[0423:TEOTMH]2.0.CO;2

Ho, S. S., Urban, A. E., and Mills, R. E. (2020). Structural Variation in the Sequencing Era: Comprehensive Discovery and Integration. *Nature Reviews. Genetics*, *21*(3), 171–189. https://doi.org/10.1038/s41576-019-0180-9

Höglund, J. (2009). *Evolutionary Conservation Genetics*. Oxford University Press.

Horrocks, N. P. C., Matson, K. D., and Tieleman, B. I. (2011). Pathogen Pressure Puts Immune Defense into Perspective. *Integrative and Comparative Biology*, *51*(4), 563–576. https://doi.org/10.1093/icb/icr011

Huang, H., Chen, D., Hu, S., Wang, J., Liu, H., Hu, J., and Li, L. (2018). Molecular characterization, expression and cellular localization of *CYP17* gene during geese (*Anser cygnoides*) follicular development. *Gene*, *658*, 184–190. https://doi.org/10.1016/j.gene.2018.03.029

Hunt, R., Sauna, Z. E., Ambudkar, S. V., Gottesman, M. M., and Kimchi-Sarfaty, C. (2009). Silent (synonymous) SNPs: Should we care about them? *Methods in Molecular Biology (Clifton, N.J.)*, *578*, 23–39. https://doi.org/10.1007/978-1-60327-411-1_2

Jackson, B. C., Campos, J. L., and Zeng, K. (2015). The effects of purifying selection on patterns of genetic differentiation between Drosophila melanogaster populations. *Heredity*, *114*(2), 163–174. https://doi.org/10.1038/hdy.2014.80

Jarulis, J., Solihin, D. D., Mardiastuti, A., and Prasetyo, L. B. (2018). DNA Barcode of Seven Indonesian Hornbills Species (Aves: Bucerotidae) Based on Mitochondrial DNA Cytochrome Oxidase Subunit I. *HAYATI Journal of Biosciences*, *25*(4), Article 4. https://doi.org/10.4308/hjb.25.4.178

Jeffares, D. C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., and Sedlazeck, F. J. (2017). Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications*, *8*, 14061. https://doi.org/10.1038/ncomms14061

Jimenez, A. G., Van Brocklyn, J., Wortman, M., and Williams, J. B. (2014). Cellular Metabolic Rate Is Influenced by Life-History Traits in Tropical and Temperate Birds. *PLoS ONE*, *9*(1), e87349. https://doi.org/10.1371/journal.pone.0087349

Jung, B., and Batal, A. b. (2012). Effect of dietary nucleotide supplementation on performance and development of the gastrointestinal tract of broilers. *British Poultry Science*, *53*(1), 98–105. https://doi.org/10.1080/00071668.2012.659654

Kanehisa, M., and Sato, Y. (2020). KEGG Mapper for inferring cellular functions from protein sequences. *Protein Science: A Publication of the Protein Society*, *29*(1), 28–35. https://doi.org/10.1002/pro.3711

Kanehisa, M., Sato, Y., and Kawashima, M. (2022). KEGG mapping tools for uncovering hidden features in biological data. *Protein Science: A Publication of the Protein Society*, *31*(1), 47–53. https://doi.org/10.1002/pro.4172

Kemp, A. C., and Kemp, M. I. (1980). The biology of the Southern Ground Hornbill *Bucorvus* leadbeateri (Vigors) (Aves: Bucerotidae). *Annals of the Transvaal Museum*, *32*(4), 65–100. https://doi.org/10.10520/AJA00411752_1057

Kim, J. M., Santure, A. W., Barton, H. J., Quinn, J. L., Cole, E. F., Consortium, G. T. H., Visser, M. E., Sheldon, B. C., Groenen, M. A. M., van Oers, K., and Slate, J. (2018). A high-density SNP chip for genotyping great tit (Parus major) populations and its application to studying the genetic architecture of exploration behaviour. *Molecular Ecology Resources*, *18*(4), 877–891. https://doi.org/10.1111/1755-0998.12778

Kraus, R. H., Kerstens, H. H., Van Hooft, P., Crooijmans, R. P., Van Der Poel, J. J., Elmberg, J., Vignal, A., Huang, Y., Li, N., Prins, H. H., and Groenen, M. A. (2011). Genome wide SNP discovery, analysis and evaluation in mallard (Anas platyrhynchos). *BMC Genomics*, *12*(1), 150. https://doi.org/10.1186/1471-2164-12-150

Lee, M. J., and Yaffe, M. B. (2016). Protein Regulation in Signal Transduction. *Cold Spring Harbor Perspectives in Biology*, *8*(6), a005918. https://doi.org/10.1101/cshperspect.a005918

Lee, S. H., and Dominguez, R. (2010). Regulation of Actin Cytoskeleton Dynamics in Cells. *Molecules and Cells*, *29*(4), 311–325.

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM* (arXiv:1303.3997). arXiv. https://doi.org/10.48550/arXiv.1303.3997

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, K. B. (2003). ClustalW-MPI: ClustalW analysis using distributed and parallel computing. *Bioinformatics (Oxford, England)*, *19*(12), 1585–1586. https://doi.org/10.1093/bioinformatics/btg192

Lin, J., Jia, P., Wang, S., Kosters, W., and Ye, K. (2023). Comparison and benchmark of structural variants detected from long read and long-read assembly. *Briefings in Bioinformatics*, *24*(4), bbad188. https://doi.org/10.1093/bib/bbad188

Lin, J., Wang, S., Audano, P. A., Meng, D., Flores, J. I., Kosters, W., Yang, X., Jia, P., Marschall, T., Beck, C. R., and Ye, K. (2022). SVision: A deep learning approach to resolve complex structural variants. *Nature Methods*, *19*(10), 1230–1233. https://doi.org/10.1038/s41592-022-01609-w

Liu, C., Ran, X., Yu, C., Xu, Q., Niu, X., Zhao, P., and Wang, J. (2019). Whole-genome analysis of structural variations between Xiang pigs with larger litter sizes and those with smaller litter sizes. *Genomics*, *111*(3), 310–319. https://doi.org/10.1016/j.ygeno.2018.02.005

Liu, T. X., Howlett, N. G., Deng, M., Langenau, D. M., Hsu, K., Rhodes, J., Kanki, J. P., D'Andrea, A. D., and Look, A. T. (2003). Knockdown of Zebrafish Fancd2 Causes Developmental Abnormalities via p53-Dependent Apoptosis. *Developmental Cell*, *5*(6), 903–914. https://doi.org/10.1016/S1534-5807(03)00339-3

Liu, Z., Roberts, R., Mercer, T. R., Xu, J., Sedlazeck, F. J., and Tong, W. (2022). Towards accurate and reliable resolution of structural variants for clinical diagnosis. *Genome Biology*, *23*, 68. https://doi.org/10.1186/s13059-022-02636-8

Lundqvist, M. L., Middleton, D. L., Radford, C., Warr, G. W., and Magor, K. E. (2006). Immunoglobulins of the non-galliform birds: Antibody expression and repertoire in the duck. *Developmental and Comparative Immunology*, *30*(1), 93–100. https://doi.org/10.1016/j.dci.2005.06.019

Ma, S. X., Seo, B. A., Kim, D., Xiong, Y., Kwon, S. H., Brahmachari, S., Kim, S., Kam, T. I., Nirujogi, R. S., Kwon, S. H., Dawson, V. L., Dawson, T. M., Pandey, A., Na, C. H., and Ko, H. S. (2021). Complement and Coagulation Cascades are Potentially Involved in Dopaminergic Neurodegeneration in α-Synuclein-Based Mouse Models of Parkinson's Disease. *Journal of Proteome Research*, *20*(7), 3428–3443. https://doi.org/10.1021/acs.jproteome.0c01002

Mabula Ground-Hornbill Project. (2024). *Mabula Ground-Hornbill Project—All you need to know*. https://ground-hornbill.org.za/about-ground-hornbills

Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C., and Sedlazeck, F. J. (2019). Structural variant calling: The long and the short of it. *Genome Biology*, *20*(1), 246. https://doi.org/10.1186/s13059-019-1828-7

Maia, R. T., and Campos, M. D. A. (2021). *Genetic Variation*. BoD – Books on Demand.

Marshall, J. S., Warrington, R., Watson, W., and Kim, H. L. (2018). An introduction to immunology and immunopathology. *Allergy, Asthma and Clinical Immunology*, *14*(2), 49. https://doi.org/10.1186/s13223-018-0278-1

Mayr, E. (1982). *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Harvard University Press.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Micheli, V., Camici, M., Tozzi, M. G., Ipata, P. L., Sestini, S., Bertelli, M., and Pompucci, G. (2011). Neurological disorders of purine and pyrimidine metabolism. *Current Topics in Medicinal Chemistry*, *11*(8), 923–947. https://doi.org/10.2174/156802611795347645

Mikkola, S. (2020). Nucleotide Sugars in Chemistry and Biology. *Molecules*, *25*(23), 5755. https://doi.org/10.3390/molecules25235755

Miller, H. C., and Lambert, D. M. (2004). Gene duplication and gene conversion in class II MHC genes of New Zealand robins (Petroicidae). *Immunogenetics*, *56*(3), 178–191. https://doi.org/10.1007/s00251-004-0666-1

Minias, P., Peng, W. X. V. H., and Matson, K. D. (2023). Evolutionary trade-off between innate and acquired immune defences in birds. *Frontiers in Zoology*, *20*(1), 32. https://doi.org/10.1186/s12983-023-00511-1

Minias, P., Pikus, E., Whittingham, L. A., and Dunn, P. O. (2018). A global analysis of selection at the avian MHC. *Evolution; International Journal of Organic Evolution*, *72*(6), 1278–1293. https://doi.org/10.1111/evo.13490

Minias, P., Włodarczyk, R., Remisiewicz, M., Cobzaru, I., and Janiszewski, T. (2021). Distinct evolutionary trajectories of MHC class I and class II genes in Old World finches and buntings. *Heredity*, *126*(6), 974–990. https://doi.org/10.1038/s41437-021-00427-8

Mitchell, P. J., Yates, R., and Parkington, J. E. (1996). At the Transition. In L. G. Straus, B. V. Eriksen, J. M. Erlandson, and D. R. Yesner (Eds.), *Humans at the End of the Ice Age: The Archaeology of the Pleistocene—Holocene Transition* (pp. 15–41). Springer US. https://doi.org/10.1007/978-1-4613-1145-4_2

Montañés, J. C., Huertas, M., Messeguer, X., and Albà, M. M. (2023). Evolutionary Trajectories of New Duplicated and Putative De Novo Genes. *Molecular Biology and Evolution*, *40*(5), msad098. https://doi.org/10.1093/molbev/msad098

Moon, B., Yang, S., Moon, H., Lee, J., and Park, D. (2023). After cell death: The molecular machinery of efferocytosis. *Experimental and Molecular Medicine*, *55*(8), 1644–1651. https://doi.org/10.1038/s12276-023-01070-5

Nanda, I., Buwe, A., Wizenman, A., Takata, M., Haaf, T., Schartl, M., and Schmid, M. (2007). Fanconi Anemia Genes in Vertebrates: Evolutionary Conservation, Sex-Linkage, and Embryonic Expression of FANCC and FANCG in Avian Cells. In D. Schindler and H. Hoehn (Eds.), *Monographs in Human Genetics* (pp. 183–199). KARGER. https://doi.org/10.1159/000102556

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, *3*(5), 418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Ng, P. C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annual Review of Genomics and Human Genetics*, *7*, 61–80. https://doi.org/10.1146/annurev.genom.7.080505.115630

Norin, T., and Metcalfe, N. B. (2019). Ecological and evolutionary consequences of metabolic rate plasticity in response to environmental change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *374*(1768), 20180180. https://doi.org/10.1098/rstb.2018.0180

Olson, N. D., Lund, S. P., Colman, R. E., Foster, J. T., Sahl, J. W., Schupp, J. M., Keim, P., Morrow, J. B., Salit, M. L., and Zook, J. M. (2015). Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics*, *6*, 235. https://doi.org/10.3389/fgene.2015.00235

Podlesak, D. W., and McWilliams, S. R. (2006). Metabolic Routing of Dietary Nutrients in Birds: Effects of Diet Quality and Macronutrient Composition Revealed Using Stable Isotopes. *Physiological and Biochemical Zoology*, *79*(3), 534–549. https://doi.org/10.1086/502813

Rada, B. (2019). Neutrophil Extracellular Traps. *Methods in Molecular Biology (Clifton, N.J.)*, *1982*, 517–528. https://doi.org/10.1007/978-1-4939-9424-3_31

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V., and Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Ruigrok, M., Xue, B., Catanach, A., Zhang, M., Jesson, L., Davy, M., and Wellenreuther, M. (2022). The Relative Power of Structural Genomic Variation versus SNPs in Explaining the Quantitative Trait Growth in the Marine Teleost Chrysophrys auratus. *Genes*, *13*(7), 1129. https://doi.org/10.3390/genes13071129

Saccheri, I., Kuussaari, M., Kankare, M., Vikman, P., Fortelius, W., and Hanski, I. (1998). Inbreeding and extinction in a butterfly metapopulation. *Nature*, *392*(6675), Article 6675. https://doi.org/10.1038/33136

Salah, M., Suprijatna, E., Djauhari, M. L., and Dwi, Y. B. I. V. (2019). The effects of nucleotide supplementation on the productivity, immune response and meat quality of broiler chicken reared under different environmental conditions. *Livestock Research for Rural Development*, *31*, 174.

Sammler, S., Havenstein, K., and Tiedemann, R. (2012). Fourteen new microsatellite markers for the Visayan tarictic hornbill (Penelopides panini) and their cross-species applicability among other endangered Philippine hornbills. *Conservation Genetics Resources*, *4*(2), 435–438. https://doi.org/10.1007/s12686-011-9567-4

Sammler, S., Ketmaier, V., Havenstein, K., Krause, U., Curio, E., and Tiedemann, R. (2012). Mitochondrial control region I and microsatellite analyses of endangered Philippine hornbill species (Aves; Bucerotidae) detect gene flow between island populations and genetic diversity loss. *BMC Evolutionary Biology*, *12*(1), 203. https://doi.org/10.1186/1471-2148-12-203

Sammler, S., Ketmaier, V., Havenstein, K., and Tiedemann, R. (2013). Intraspecific Rearrangement of Duplicated Mitochondrial Control Regions in the Luzon Tarictic Hornbill Penelopides manillae (Aves: Bucerotidae). *Journal of Molecular Evolution*, *77*(5), 199–205. https://doi.org/10.1007/s00239-013-9591-y

Satyam, A., Graef, E. R., Lapchak, P. H., Tsokos, M. G., Dalle Lucca, J. J., and Tsokos, G. C. (2019). Complement and coagulation cascades in trauma. *Acute Medicine and Surgery*, *6*(4), 329–335. https://doi.org/10.1002/ams2.426

Schlinger, B. A. (2015). Steroids in the Avian Brain: Heterogeneity across Space and Time. *Journal of Ornithology / DO-G*, *156*(Suppl 1), 419–424. https://doi.org/10.1007/s10336-015-1184-7

Schlinger, B. A., Lane, N. I., Grisham, W., and Thompson, L. (1999). Androgen Synthesis in a Songbird: A Study of Cyp17 (17α-Hydroxylase/C17,20-Lyase) Activity in the Zebra

Finch. *General and Comparative Endocrinology*, *113*(1), 46–58. https://doi.org/10.1006/gcen.1998.7179

Schoville, S. D., Bonin, A., François, O., Lobreaux, S., Melodelima, C., and Manel, S. (2012). Adaptive Genetic Variation on the Landscape: Methods and Cases. *Annual Review of Ecology, Evolution, and Systematics*, *43*(1), 23–43. https://doi.org/10.1146/annurev-ecolsys-110411-160248

Silva, V. H. da. (2020). *Structural variants in the great tit genome and their effect on seasonal timing*. https://doi.org/10.18174/512907

Smolka, M., Paulin, L. F., Grochowski, C. M., Horner, D. W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S. W., Carvalho, C. M. B., Proukakis, C., and Sedlazeck, F. J. (2024). Detection of mosaic and population-level structural variants with Sniffles2. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-023-02024-y

Steinway, S. N., Dannenfelser, R., Laucius, C. D., Hayes, J. E., and Nayak, S. (2010). JCoDA: A tool for detecting evolutionary selection. *BMC Bioinformatics*, *11*(1), 284. https://doi.org/10.1186/1471-2105-11-284

Strillacci, M. G., Cozzi, M. C., Gorla, E., Mosca, F., Schiavini, F., Román-Ponce, S. I., Ruiz López, F. J., Schiavone, A., Marzoni, M., Cerolini, S., and Bagnato, A. (2017). Genomic and genetic variability of six chicken populations using single nucleotide polymorphism and copy number variants as markers. *Animal*, *11*(5), 737–745. https://doi.org/10.1017/S1751731116002135

Swanson, D. L., Stager, M., Vézina, F., Liu, J. S., McKechnie, A. E., and Amirkhiz, R. G. (2023). Evidence for a maintenance cost for birds maintaining highly flexible basal, but not summit, metabolic rates. *Scientific Reports*, *13*(1), 8968. https://doi.org/10.1038/s41598-023-36218-w

Tan, X., Zhang, J., Dong, J., Huang, M., Li, Q., Wang, H., Bai, L., Cui, M., Zhou, Z., Yang, S., and Wang, D. (2024). Whole-genome variants dataset of 209 local chickens from China. *Scientific Data*, *11*(1), 169. https://doi.org/10.1038/s41597-024-02995-w

Theron, N., Dalton, D., Grobler, J. P., Jansen, R., and Kotze, A. (2013). Molecular insights on the re-colonization of the Limpopo Valley, South Africa, by Southern Ground-Hornbills. *Journal of Ornithology*, *154*(3), Article 3. https://doi.org/10.1007/s10336-013-0937-4

Tine, M., Kuhl, H., Teske, P. R., Tschöp, M. H., and Jastroch, M. (2016). Diversification and coevolution of the ghrelin/growth hormone secretagogue receptor system in vertebrates. *Ecology and Evolution*, *6*(8), 2516–2535. https://doi.org/10.1002/ece3.2057

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K. V., Altshuler, D., Gabriel, S., and DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et al.]*, *11*(1110), 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

van Perlo. (1999). *Birds of Southern Africa*. HarperCollins. https://typeset.io/papers/birds-of-southern-africa-n9crfyd9eg

Vihinen, M. (2015). Types and effects of protein variations. *Human Genetics*, *134*(4), 405–421. https://doi.org/10.1007/s00439-015-1529-6

Viseshakul, N., Charoennitikul, W., Kitamura, S., Kemp, A., Thong-Aree, S., Surapunpitak, Y., Poonswad, P., and Ponglikitmongkol, M. (2011). A phylogeny of frugivorous hornbills linked to the evolution of Indian plants within Asian rainforests. *Journal of Evolutionary Biology*, *24*(7), 1533–1545. https://doi.org/10.1111/j.1420-9101.2011.02285.x

Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K. J., Knief, U., Heim, W., Peona, V., Pophaly, S. D., Sedlazeck, F. J., Suh, A., Warmuth, V. M., and Wolf, J. B. W. (2020). Discovery and population genomics of structural variation in a songbird genus. *Nature Communications*, *11*(1), 3403. https://doi.org/10.1038/s41467-020-17195-4

White, C. R., Blackburn, T. M., Martin, G. R., and Butler, P. J. (2007). Basal metabolic rate of birds is associated with habitat temperature and precipitation, not primary productivity. *Proceedings of the Royal Society B: Biological Sciences*, *274*(1607), 287–293. https://doi.org/10.1098/rspb.2006.3727

Wold, J. R., Guhlin, J. G., Dearden, P. K., Santure, A. W., and Steeves, T. E. (2023). The promise and challenges of characterizing genome-wide structural variants: A case study in a critically endangered parrot. *Molecular Ecology Resources*, *n/a*(n/a). https://doi.org/10.1111/1755-0998.13783

Wone, B. W. M., and Swanson, D. L. (2022). Metabolic Profiling and Integration of Metabolomic and Transcriptomic Data From Pectoralis Muscle Reveal Winter-Adaptive Metabolic Responses of Black-Capped Chickadee and American Goldfinch. *Frontiers in Ecology and Evolution*, *10*. https://doi.org/10.3389/fevo.2022.866130

Wu, L., Jiao, X., Zhang, D., Cheng, Y., Song, G., Qu, Y., and Lei, F. (2021). Comparative Genomics and Evolution of Avian Specialized Traits. *Current Genomics*, *22*(7), 496–511. https://doi.org/10.2174/1389202923666211227143952

Yang, J., Xie, D., and Ma, X. (2023). Recent Advances in Chemical Synthesis of Amino Sugars. *Molecules*, *28*(12), 4724. https://doi.org/10.3390/molecules28124724

Yang, Z., and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. *Molecular Biology and Evolution*, *17*(1), 32–43. https://doi.org/10.1093/oxfordjournals.molbev.a026236

Yin, H., Ma, H., Hussain, S., Zhang, H., Xie, X., Jiang, L., Jiang, X., Iqbal, F., Bukhari, I., Jiang, H., Ali, A., Zhong, L., Li, T., Fan, S., Zhang, B., Gao, J., Li, Y., Nazish, J., Khan, T., … Shi, Q. (2019). A homozygous FANCM frameshift pathogenic variant causes male infertility. *Genetics in Medicine*, *21*(1), 62–70. https://doi.org/10.1038/s41436-018-0015-7

Yusuf, L., Heatley, M. C., Palmer, J. P. G., Barton, H. J., Cooney, C. R., and Gossmann, T. I. (2020). Noncoding regions underpin avian bill shape diversification at macroevolutionary scales. *Genome Research*, *30*(4), 553–565. https://doi.org/10.1101/gr.255752.119

Zhang, J., Nie, C., Li, X., Zhao, X., Jia, Y., Han, J., Chen, Y., Wang, L., Lv, X., Yang, W., Li, K., Zhang, J., Ning, Z., Bao, H., Zhao, C., Li, J., and Qu, L. (2022). Comprehensive analysis of structural variants in chickens using PacBio sequencing. *Frontiers in Genetics*, *13*, 971588. https://doi.org/10.3389/fgene.2022.971588

Zhang, X., Li, Z., Liu, W., Du, J., Liu, Y., Yu, N., Liu, C., Zeng, M., and Zhang, X. (2022). The Complement and Coagulation Cascades Pathway is Associated with Acute Necrotizing Pancreatitis by Genomics and Proteomics Analysis. *Journal of Inflammation Research*, *15*, 2349–2363. https://doi.org/10.2147/JIR.S351416

Zhu, F., Cui, Q. Q., and Hou, Z. C. (2016). SNP discovery and genotyping using Genotyping-by-Sequencing in Pekin ducks. *Scientific Reports*, *6*, 36223. https://doi.org/10.1038/srep36223

SUMMARY

The Southern Ground Hornbill (*Bucorvus leadbeateri*) is under threat, due to declining populations and loss of their natural habitat. As part of the current conservation efforts, we have sequenced the complete genome of the SGH and performed comprehensive comparative genomic analyses with other hornbill species.

The genome of the SGH was sequenced using a combination of long- and short-read sequencing technologies and assembled to high-quality draft status using a refined hybrid *de novo* assembly approach. The final genome comprises 1,672 contigs (N50 value of 40.45 Mb) with a total size of 1.16 Gb.

Comparative genomic analysis revealed that the genome size of the *Bucorvus* species (SGH and AGH) was 60.52 Mb larger than that of two *Buceros* species (GH and RH). Additionally, they coded for 30.31% more proteins than those of the latter genus. These variations could be attributed to differences in body size (ground hornbills are substantially larger than all hornbill species), geographical distribution (Africa vs Asia) and lifestyle patterns (terrestrial vs arboreal) of these birds. Functional analysis of the accessory proteome shared between the *Bucorvus* species and the *Buceros* species further highlighted differences in the abundance of proteins associated with metabolism and a range of cellular processes. The disparate proteomes may have set members of the two genera on their distinct evolutionary trajectories, allowing their adaptation to their respective environments.

Analysis of the proteomes of the SGH and AGH also highlighted extensive differences between their genomes, in particular proteins associated with metabolism and immune response. Moreover, 51,779 genetic variants with a high impact on protein-coding sequences distinguish the SGH and AGH genomes, several of which are predicted to play a role in immune response via MHC class II antigens. Selection analysis further confirmed proteins involved in reproduction were under negative selection, which could have crucial implications for the survival and fitness of these bird species. The SGH has a complex repertoire of species-specific and gene-duplicated proteins that varies from the AGH which may have provided it with a selective advantage in adapting to its habitat south of the equator in the African savannah. Furthermore, it has maintained a delicate balance between essential biological functions and environmental adaptations. Ultimately, this study creates a pivotal foundation for large-scale

population genetic analysis of the SGH which will aid in conservation efforts to help save the remaining population of this exceptional bird species.

APPENDIX

Appendix Table 1: Impact of changes in parameters of the Sparse Assembler and overlap and layout step of the DBG2OLC assembly

Assembly	Number of contigs	Largest contig (bp)	Total genome length (bp)	GC (%)	N50 (bp)	L50 (bp)	Genome fraction (%) [1]
k-mer length							
15	2,887,879	27,799	1,158,960,294	43.48	2,118	153,161	90.02
21	2,887,879	27,799	1,158,960,294	43.48	2,118	153,161	90.02
31	2,887,879	27,799	1,158,960,294	43.48	2,118	153,161	90.02
51	2,887,879	27,799	1,158,960,294	43.48	2,118	153,161	90.02
71	6,326,767	15,132	1,423,653,792	43.77	326	1,208,448	78.68
Gap size							
1	2,887,879	27,799	1,158,960,294	43.48	2,118	153,161	90.02
5	2,349,503	27,971	1,134,048,907	43.32	2,279	140,491	90.06
7	2,002,317	27,882	1,117,722,795	43.23	2,420	131,222	90.14
10	1,668,676	28,536	1,100,601,241	43.13	2,589	121,749	90.27
11	1,593,430	31,906	1,094,237,748	43.10	2,587	121,454	90.12
12	1,509,338	27,792	1,090,826,962	43.07	2,688	116,701	90.25
13	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
14	1,402,831	28,541	1,083,721,253	43.04	2,731	114,492	90.18
15	1,346,379	35,070	1,081,195,482	43.02	2,796	111,773	90.25
25	1,057,994	34,473	1,058,680,414	42.91	2,905	106,063	89.74
NodeCovTh							
0	6,470,671	18,762	1,240,972,141	43.22	566	548,444	80.30
1	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
2	2,005,420	17,043	1,047,076,024	42.97	1,292	235,093	84.46
5	4,708,924	19,125	809,532,014	43.00	251	940,441	41.61
10	2,987,995	16,885	248,495,110	43.57	96	848,954	1.94
EdgeCovTh							
0	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
1	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
2	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
5	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
10	1,448,085	27,842	1,087,371,920	43.05	2,735	114,623	90.30
AdaptiveTh							
0.0001	10,639	1,493,494	1,112,291,772	43.50	245,809	1,346	95.13
0.001	10,536	2,059,266	1,114,952,537	43.49	249,633	1,316	95.40

0.002	10,479	2,059,072	1,116,085,317	43.50	257,244	1,291	95.46
0.003	10,543	2,059,065	1,116,533,394	43.50	258,977	1,287	95.46
0.004	10,585	1,930,759	1,116,669,945	43.51	258,267	1,285	95.46
0.005	10,672	1,930,808	1,116,790,880	43.51	257,065	1,285	95.45
0.006	10,726	1,496,268	1,116,789,610	43.51	256,183	1,302	95.43
0.01	11,058	1,496,274	1,114,781,605	43.51	244,249	1,367	95.29
0.02	11,949	1,343,768	1,109,100,481	43.51	214,063	1,548	94.80
KmerCov							
2	10,543	2,059,065	1,116,533,394	43.50	258,977	1,287	95.46
3	10,782	2,073,508	1,118,582,400	43.51	256,655	1,304	95.46
5	11,055	1,558,216	1,120,337,309	43.52	251,329	1,330	95.42
MinOverlap							
19	22,700	745,460	1,093,101,638	44.30	90,584	3,557	80.72
20	10,543	2,059,065	1,116,533,394	43.50	258,977	1,287	95.46
21	10,518	2,058,996	1,116,083,548	43.50	258,489	1,289	95.47
22	10,485	2,059,011	1,115,298,025	43.49	258,047	1,293	95.43

[1] Genome fraction calculated based on the proportion of the SGH genome that mapped against the reference genome of the AGH which had a genome length of 1,132,614,622 bp.

Appendix Table 2: Performance evaluation of draft genome assembly improvement tools

Assembly	Number of contigs (bp)	Largest contig	Total genome length (bp)	GC (%)	N50 (bp)	L50 (bp)	Genome fraction (%)[1]	Number of N's per 100 kb
Scaffolding								
SSPACE_LR								
HiFi reads	10,491	2,058,996	1,116,024,139	43.50	258,489	1,289	-	0.00
CLR reads	10,491	2,058,996	1,116024139	43.50	258,489	1,289	-	0.00
SSPACE-Basic								
Default	10,516	2,058,996	1,116,136,816	43.50	258,489	1,289	95.47	0.00
Optimized	10,513	2,058,996	1,116,441,123	43.50	258,626	1,289	95.50	0.00
Polishing								
Racon	10,518	2,025,051	1,098,548,457	43.49	254,827	1,288	95.22	0.00
Pilon								
1st	10,513	2,058,352	1,115,980,173	43.50	258,554	1,289	95.50	0.00
2nd	10,513	2,058,338	1,115,942,927	43.50	258,551	1,289	95.50	0.00
3rd	10,513	2,058,333	1,115,919,892	43.50	258,551	1,289	95.50	0.00
4th	10,513	2,058,332	1,115,907,610	43.50	258,551	1,289	95.50	0.00
5th	10,513	2,058,332	1,115,898,967	43.50	258,551	1,289	95.50	0.00
6th	10,513	2,058,332	1,115,896,358	43.50	258,551	1,289	95.50	0.00
7th	10,513	2,058,332	1,115,893,644	43.50	258,551	1,289	95.50	0.00
8th	10,513	2,058,332	1,115,887,539	43.50	258,551	1,289	95.50	0.00
9th	10,513	2,058,332	1,115,879,654	43.50	258,551	1,289	95.50	0.00
10th	10,513	2,058,332	1,115,874,943	43.50	258,551	1,289	95.50	0.00
Scaffolding								
Chromosome Scaffolder	1,808	134,263,929	1,122,420,046	43.50	38,661,887	10	95.09	590.87
Ragtag	1,653	138,051,206	1,160,672,592	43.50	40,478,097	9	95.18	3863.56
Long-read gap closing								
TGS-GapCloser								
HiFi reads	1,653	137,935,013	1,158,765,572	43.50	40,411,018	9	95.22	3688.41
CLR reads	1,653	137,837,006	1,158,326,194	43.50	40,410,604	9	95.24	3650.90
LR_GapCloser								
HiFi reads	1,653	138,052,830	1,160,685,000	43.50	40,478,587	9	95.66	3363.74
CLR reads	1,653	138,055,092	1,160,708,975	43.50	40,479,380	9	95.84	3172.10
Combination	1,653	138,053,805	1,160,701,909	43.51	40,478,783	9	95.85	3157.46
Samba								
Combination	1,653	138,051,206	1,160,673,239	43.50	40,478,097	9	95.18	3863.56
Short-read gap closer								

GapCloser	1,674	138,052,837	1,160,719,150	43.51	40,479,000	9	96.30	2707.69
Gap2Seq	1,674	138,053,805	1,160,747,019	43.51	40,478,646	9	95.85	3157.30
Sealer	1,674	138,051,659	1,160,726,886	43.51	40,478,530	9	95.85	3150.53
GapCloser and Sealer combination	1,674	138,051,936	1,160,714,302	43.51	40,478,974	9	96.31	2703.41
Polishing								
Pilon								
1st	1,653	138,034,778	1,160,196,490	43.51	40,467,238	9	96.28	2703.58
2nd	1,653	138,034,777	1,159,870,375	43.51	40,463,888	9	96.26	2704.01
3rd	1,653	138,034,742	1,159,674,072	43.51	40,463,892	9	96.25	2704.33
4th	1,653	138,034,743	1,159,589,526	43.51	40,463,885	9	96.24	2704.48
5th	1,653	138,034,744	1,159,524,602	43.51	40,463,886	9	96.24	2704.57
PolCa								
1st	1,653	137,991,452	1,159,142,644	43.50	40,451,646	9	96.28	2705.30
2nd	1,653	137,990,668	1,159,133,141	43.50	40,451,622	9	96.29	2705.17
3rd	1,653	137,990,526	1,159,131,223	43.50	40,451,526	9	96.29	2705.03
4th	1,653	137,990,495	1,159,130,546	43.50	40,451,540	9	96.29	2704.90
5th	1,653	137,990,494	1,159,130,246	43.50	40,451,535	9	96.29	2704.78

[1] Genome fraction calculated based on the proportion of the SGH genome that mapped against the reference genome of the AGH which had a genome length of 1,132,614,622 bp.

Appendix Table 3: Summary of the number of changes made by Pilon during the first and second polishing steps

	First step of polishing	Second step of polishing
1st iteration	1,060,618	49,795
2nd iteration	55,092	7,720
3rd iteration	10,985	2,232
4th iteration	3,592	1,078
5th iteration	1,797	612
6th iteration	992	
7th iteration	647	
8th iteration	476	
9th iteration	396	
10th iteration	323	

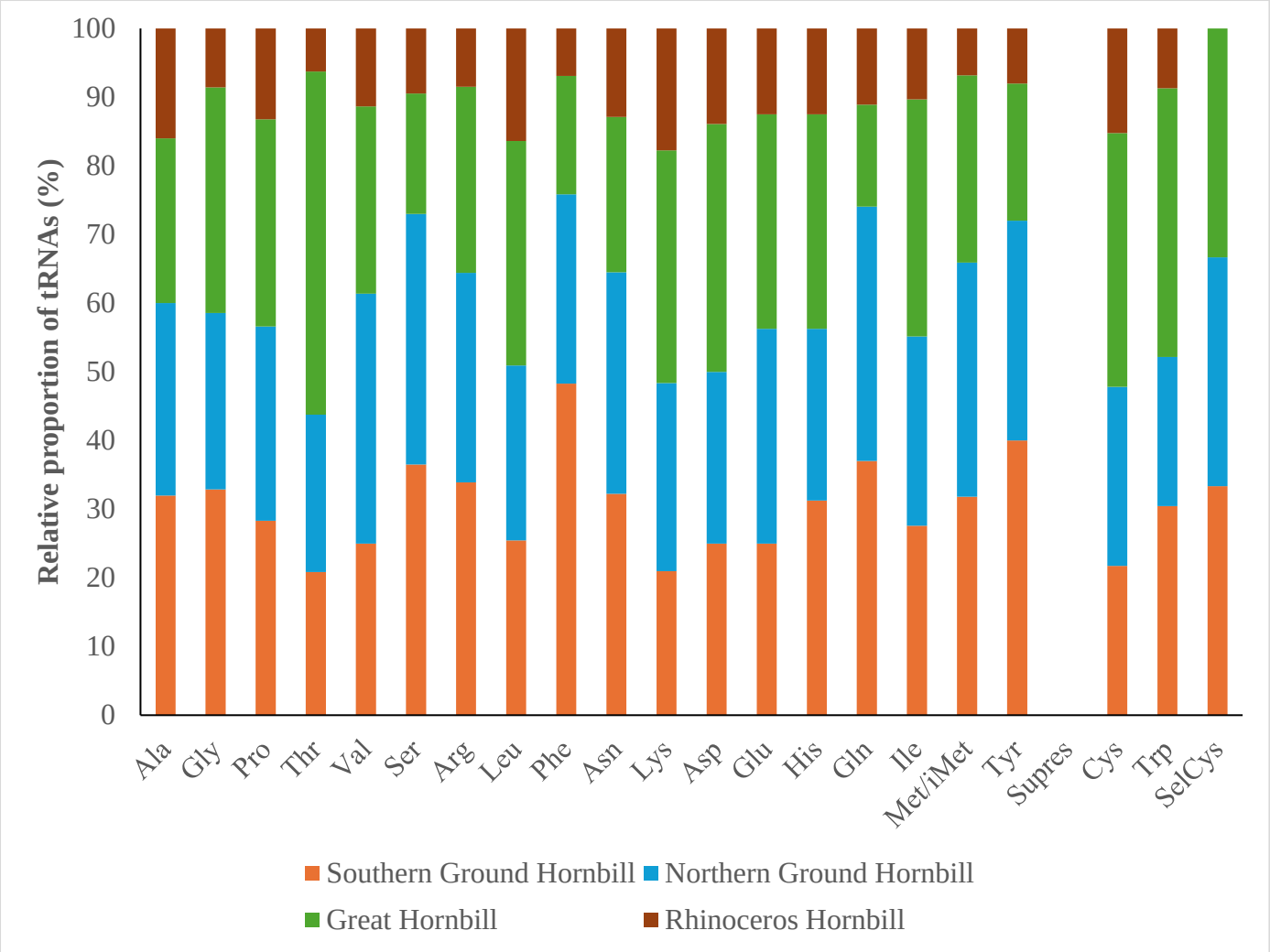
Appendix Table 4: Summary of the changes made by PolCa over five iterations of polishing

	Substitutions	**INDELs**
1st iteration	35,240	449,888
2nd iteration	7,002	61,989
3rd iteration	2,832	34,286
4th iteration	1,536	27,937
5th iteration	1,151	23,668

Appendix Table 5: Sizes of the SGH scaffolds that aligned with the AGH chromosomes

SGH locus tag	Gapped size (bp)	Ungapped size (bp)	AGH locus tag	Chromosome number	Gapped size (bp)	Ungapped size (bp)
Scaff00001_Aln_CM020024.1	137,990,492	135,563,958	CM020024.1	Chromosome 1	136,452,362	135,661,717
Scaff00002_Aln_CM020025.1	80,918,672	79,767,627	CM020025.1	Chromosome 2	81,184,632	80,561,190
Scaff00003_Aln_CM020026.1	63,345,211	62,632,314	CM020026.1	Chromosome 3	62,186,786	62,049,764
Scaff00004_Aln_CM020027.1	47,442,010	47,114,594	CM020027.1	Chromosome 4	46,835,259	46,798,603
Scaff00005_Aln_CM020028.1	46,096,250	45,609,770	CM020028.1	Chromosome 5	45,440,818	45,315,996
Scaff00006_Aln_CM020029.1	42,218,628	41,882,859	CM020029.1	Chromosome 6	41,866,552	41,775,197
Scaff00007_Aln_CM020030.1	41,680,333	41,086,556	CM020030.1	Chromosome 7	41,247,383	41,225,079
Scaff00008_Aln_CM020031.1	40,451,533	39,825,503	CM020031.1	Chromosome 8	40,274,051	40,077,076
Scaff00009_Aln_CM020032.1	39,217,941	38,935,356	CM020032.1	Chromosome 9	39,094,385	39,094,272
Scaff00010_Aln_CM020033.1	34,685,309	34,257,489	CM020033.1	Chromosome 10	35,074,608	34,583,428
Scaff00011_Aln_CM020034.1	34,255,575	33,989,563	CM020034.1	Chromosome 11	34,097,363	34,037,142
Scaff00012_Aln_CM020035.1	27,228,310	26,960,054	CM020035.1	Chromosome 12	27,197,564	27,049,715
Scaff00013_Aln_CM020036.1	25,326,201	25,058,158	CM020036.1	Chromosome 13	25,328,146	25,328,146
Scaff00014_Aln_CM020037.1	23,159,027	23,037,031	CM020037.1	Chromosome 14	22,999,164	22,998,551
Scaff00015_Aln_CM020038.1	22,554,558	22,399,588	CM020038.1	Chromosome 15	22,535,586	22,535,286
Scaff00016_Aln_CM020039.1	22,600,731	22,477,871	CM020039.1	Chromosome 16	22,493,967	22,461,494
Scaff00017_Aln_CM020040.1	21,279,581	21,143,356	CM020040.1	Chromosome 17	22,365,827	20,837,535
Scaff00018_Aln_CM020041.1	20,707,880	20,494,726	CM020041.1	Chromosome 18	20,536,565	20,536,565
Scaff00019_Aln_CM020042.1	20,810,109	20,659,684	CM020042.1	Chromosome 19	20,321,109	20,320,509
Scaff00020_Aln_CM020043.1	18,881,779	18,652,354	CM020043.1	Chromosome 20	19,022,054	18,617,941
Scaff00021_Aln_CM020044.1	18,206,627	18,066,987	CM020044.1	Chromosome 21	18,028,713	18,028,513
Scaff00022_Aln_CM020045.1	17,753,702	17,607,295	CM020045.1	Chromosome 22	17,599,546	17,525,647

Scaff00023_Aln_CM020046.1	17,718,755	17,565,124	CM020046.1	Chromosome 23	17,575,511	17,539,150
Scaff00024_Aln_CM020047.1	17,068,996	16,985,952	CM020047.1	Chromosome 24	16,590,994	16,357,641
Scaff00025_Aln_CM020048.1	15,693,606	15,594,492	CM020048.1	Chromosome 25	15,599,372	15,554,905
Scaff00026_Aln_CM020049.1	14,590,094	14,339,181	CM020049.1	Chromosome 26	15,563,623	14,544,968
Scaff00027_Aln_CM020050.1	14,996,587	14,684,896	CM020050.1	Chromosome 27	14,657,043	14,559,447
Scaff00028_Aln_CM020051.1	13,444,535	13,337,785	CM020051.1	Chromosome 28	13,341,052	13,340,952
Scaff00029_Aln_CM020052.1	13,355,021	13,221,347	CM020052.1	Chromosome 29	13,216,842	13,216,342
Scaff00030_Aln_CM020053.1	13,242,328	13,073,645	CM020053.1	Chromosome 30	13,142,313	13,142,113
Scaff00031_Aln_CM020054.1	13,418,007	13,315,510	CM020054.1	Chromosome 31	13,116,243	13,116,143
Scaff00032_Aln_CM020055.1	12,895,310	12,541,693	CM020055.1	Chromosome 32	12,593,932	12,498,158
Scaff00033_Aln_CM020056.1	12,302,255	12,247,806	CM020056.1	Chromosome 33	12,224,544	12,224,444
Scaff00034_Aln_CM020057.1	7,678,295	7,595,588	CM020057.1	Chromosome 34	7,491,012	7,403,783
Scaff00035_Aln_CM020058.1	7,269,654	7,111,637	CM020058.1	Chromosome 35	7,153,002	7,136,819
Scaff00036_Aln_CM020059.1	5,240,864	5,105,445	CM020059.1	Chromosome 36	5,208,711	5,111,518
Scaff00037_Aln_CM020060.1	3,627,919	3,490,358	CM020060.1	Chromosome 37	3,539,624	3,538,724
Scaff00038_Aln_CM020061.1	3,007,457	2,871,617	CM020061.1	Chromosome 38	2,976,411	2,975,711
Scaff00039_Aln_CM020062.1	1,318,437	988,124	CM020062.1	Chromosome 39	1,271,767	1,268,967
Scaff00040_Aln_CM020063.1	81,607,035	68,471,498	CM020063.1	Chromosome Z	82,367,667	81,228,392
Scaff00041_Aln_CM020064.1	15,929,420	10,816,632	CM020064.1	Chromosome W	16,450,271	16,019,189



Appendix Figure 1: tRNA isotype abundance within the hornbill and chicken genomes. tRNA isotypes are depicted in the three-letter amino acid code and the Supres label indicates the tRNA suppressor isotypes coding for stop codons.

Appendix Table 6: COG classifications of the core, unique and shared proteins identified in the hornbill species

COG category	Function	Core	Unique				Shared accessory genome	
			SGH	AGH	GH	RH	*Bucorvus*	*Buceros*
Information storage and processing								
A	RNA processing and modification	396	12	22	37	38	52	23
B	Chromatin Structure and dynamics	172	12	10	23	27	14	6
J	Translation	273	15	17	44	36	37	15
K	Transcription	1,203	86	71	139	144	173	65
L	Replication and repair	548	165	203	162	30	249	33
Cellular processes and signalling								
D	Cell cycle control and mitosis	259	27	13	23	40	19	6
M	Cell wall/membrane/envelop biogenesis	73	5	2	68	9	4	1
N	Cell motility	13	1	0	21	5	2	0
O	Post-translational modification, protein turnover, chaperone functions	1,095	136	89	125	159	245	142
T	Signal Transduction	3,151	205	158	222	533	359	167
U	Intracellular trafficking and secretion	610	32	33	66	94	51	31
V	Defence mechanism	153	17	7	13	18	13	8
W	Extracellular structures	236	11	12	10	29	20	13
X	Mobilome	0	0	0	0	0	0	0
Y	Nuclear structure	41	4	2	0	4	1	4
Z	Cytoskeleton	529	45	31	23	120	44	20
Metabolism								
C	Energy production and conversion	217	41	12	63	20	53	13
E	Amino Acid metabolism and transport	762	346	266	375	296	1,049	729
F	Nucleotide metabolism and transport	140	15	7	16	18	15	6
G	Carbohydrate metabolism and transport	434	54	31	85	57	142	86
H	Coenzyme metabolism	71	4	4	34	6	3	4
I	Lipid metabolism	368	22	14	41	43	26	12
P	Inorganic ion transport and metabolism	444	15	23	78	82	39	25
Q	Secondary Structure	175	20	3	41	35	31	22
Poorly characterized								
R	General function prediction only	0	0	0	6	0	0	0
S	Function Unknown	5,616	1,458	1,207	785	715	3,085	1,300

-	Undetermined	84	27	11	232	3	18	10
Total		**17,062**	**2,775**	**2,248**	**2,732**	**2,561**	**5,738**	**2,737**

Appendix Table 7: KEGG pathways reconstructed for the shared accessory genome and paralogous proteins in the hornbill species. Pathways exhibiting 5-fold more proteins in the *Bucorvus* species are highlighted in red and those exhibiting 5-fold more proteins in the *Buceros* species in orange.

		Shared accessory			Paralogous proteins		
KEGG class	**KEGG category**	*Bucorvus*	*Buceros*	Factor	*Bucorvus*	*Buceros*	Factor
Metabolism	Global and overview maps	115,5	60	1,93	3,5	1,5	2,33
	Carbohydrate metabolism	23	16	1,44	0	0	0,00
	Energy metabolism	20	4	5,00	0,5	0	0,50
	Lipid metabolism	26,5	3,5	7,57	1	0	1,00
	Nucleotide metabolism	8,5	4	2,13	1,5	0,5	3,00
	Amino acid metabolism	21,5	17,5	1,23	0	0	0,00
	Metabolism of other amino acids	6,5	3	2,17	0	0	0,00
	Glycan biosynthesis and metabolism	13	7,5	1,73	1	1	1,00
	Metabolism of cofactors and vitamins	6	5	1,20	0	0	0,00
	Metabolism of terpenoids and polyketides	2	0,5	4,00	0	0	0,00
	Biosynthesis of other secondary metabolites	4	5	0,80	0	0	0,00
	Xenobiotics biodegradation and metabolism	3	2	1,50	1	0,5	2,00
Genetic information processing	Transcription	13,5	4	3,38	1	1	1,00
	Translation	31	15	2,07	1,5	0,5	3,00
	Folding, sorting and degradation	28,5	15,5	1,84	1,5	0,5	3,00
	Replication and repair	17,5	6,5	2,69	1	0	1,00
	Chromosome	7,5	4	1,88	2	1	2,00
	Information processing in viruses	4	1	4,00	0,5	0	0,50
Environmental information processing	Membrane transport	1	0	0,00	0	0	0,00
	Signal transduction	186,5	82	2,27	22	11,5	1,91
	Signalling molecules and interaction	33,5	17	1,97	3	1	3,00
	Transport and catabolism	40,5	20,5	1,98	6,5	4	1,63
	Cell growth and death	40,5	21	1,93	4,5	4	1,13
	Cellular community - eukaryotes	30	20,5	1,46	2,5	0,5	5,00
	Cell motility	22,5	12,5	1,80	1,5	1,5	1,00
Organism systems	Immune system	90,5	32,5	2,78	25	15,5	1,61
	Endocrine system	82	70,5	1,16	6,5	1,5	4,33
	Circulatory system	11,5	12,5	0,92	0	0,5	0,50
	Digestive system	17,5	18	0,97	1,5	0	1,50
	Excretory system	10,5	1,5	7,00	0,5	0	0,50
	Nervous system	58	36,5	1,59	1,5	0	1,50
	Sensory system	8,5	8,5	1,00	0,5	0,5	1,00

	Development and regeneration	22	9	2,44	1	0	1,00
	Aging	12,5	5	2,50	1,5	0	1,50
	Environmental adaptation	19	7	2,71	0,5	0	0,50
Human diseases	Cancer: overview	81,5	35,5	2,30	0	0	0,00
	Cancer: specific types	55,5	27,5	2,02	0	0	0,00
	Infectious disease: viral	95,5	34	2,81	0	0	0,00
	Infectious disease: bacterial	57	21	2,71	0	0	0,00
	Infectious disease: parasitic	18,5	9,5	1,95	0	0	0,00
	Immune disease	23	14	1,64	0	0	0,00
	Neurodegenerative disease	117	64	1,83	0	0	0,00
	Substance dependence	32,5	16	2,03	0	0	0,00
	Cardiovascular disease	34	24	1,42	0	0	0,00
	Endocrine and metabolic disease	34,5	14	2,46	0	0	0,00
	Drug resistance: antineoplastic	13	4	3,25	0	0	0,00
	Total	**1600**	**812**		**94,5**	**47**	

Appendix Table 8: Shared proteins in the accessory genome involved in lipid metabolism of the hornbill species. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species

Pathway	Protein name	Locus tags	COG category	KEGG KO term
Fatty acid biosynthesis	Enoylreductase	Scaff00028_Aln_CM020051.1.g9.t1 CM020051.1.g9.t1	Q	K00665
	S-acyl fatty acid synthase thioesterase, medium chain	Scaff00005_Aln_CM020028.1.g566.t1 CM020028.1.g572.t1	Q	K01071
Fatty acid elongation	Very long chain fatty acid elongase 6	Scaff00003_Aln_CM020026.1.g1598.t1 CM020026.1.g1637.t1	I	K10203
	Thioesterase superfamily	Scaff00038_Aln_CM020061.1.g65.t1 CM020061.1.g69.t1	S	K16339
Fatty acid degradation	acyl-CoA dehydrogenase	Scaff00025_Aln_CM020048.1.g309.t1 CM020048.1.g312.t1	I	K00248
Primary bile acid biosynthesis	Peroxisomal multifunctional enzyme type 2	Scaff00040_Aln_CM020063.1.g1395.t1 CM020063.1.g1794.t1	Q	K12405
Steroid hormone biosynthesis	Corticosteroid 11-beta-dehydrogenase isozyme 2	Scaff00017_Aln_CM020040.1.g63.t1 CM020040.1.g55.t1	Q	K00071
	Steroid 5 alpha-reductase 3	JAODLK010000035.1.g656.t1	I	K12345
Glycerolipid metabolism	Aldo keto reductase family	Scaff00002_Aln_CM020025.1.g1759.t1 CM020025.1.g1839.t1	S	K00011
	glycerol kinase	Scaff00001_Aln_CM020024.1.g181.t1, Scaff00001_Aln_CM020024.1.g769.t1, Scaff00001_Aln_CM020024.1.g806.t1, Scaff00001_Aln_CM020024.1.g923.t1, Scaff00001_Aln_CM020024.1.g1167.t1, Scaff00001_Aln_CM020024.1.g1617.t1, Scaff00001_Aln_CM020024.1.g1725.t1, Scaff00001_Aln_CM020024.1.g1877.t1, Scaff00001_Aln_CM020024.1.g2122.t1, Scaff00001_Aln_CM020024.1.g2133.t1, Scaff00001_Aln_CM020024.1.g2175.t1, Scaff00001_Aln_CM020024.1.g2291.t1, Scaff00001_Aln_CM020024.1.g2402.t1, Scaff00001_Aln_CM020024.1.g2527.t1, Scaff00001_Aln_CM020024.1.g2639.t1, Scaff00001_Aln_CM020024.1.g2648.t1, Scaff00001_Aln_CM020024.1.g2665.t1, Scaff00001_Aln_CM020024.1.g2971.t1, Scaff00002_Aln_CM020025.1.g685.t1, Scaff00002_Aln_CM020025.1.g795.t1, Scaff00002_Aln_CM020025.1.g1095.t1, Scaff00002_Aln_CM020025.1.g1132.t1, Scaff00002_Aln_CM020025.1.g2053.t1, Scaff00003_Aln_CM020026.1.g684.t1, Scaff00003_Aln_CM020026.1.g709.t1, Scaff00003_Aln_CM020026.1.g1049.t1, Scaff00003_Aln_CM020026.1.g1069.t1, Scaff00003_Aln_CM020026.1.g1418.t1, Scaff00003_Aln_CM020026.1.g1664.t1, Scaff00004_Aln_CM020027.1.g17.t1,	G	K00864

		Scaff00004_Aln_CM020027.1.g422.t1, Scaff00004_Aln_CM020027.1.g445.t1, Scaff00004_Aln_CM020027.1.g649.t1, Scaff00004_Aln_CM020027.1.g707.t1, Scaff00004_Aln_CM020027.1.g948.t1, Scaff00007_Aln_CM020030.1.g320.t1,		
		Scaff00008_Aln_CM020031.1.g387.t1, Scaff00009_Aln_CM020032.1.g673.t1, Scaff00010_Aln_CM020033.1.g25.t1, Scaff00010_Aln_CM020033.1.g113.t1, Scaff00010_Aln_CM020033.1.g151.t1, Scaff00013_Aln_CM020036.1.g436.t1, Scaff00013_Aln_CM020036.1.g469.t1, Scaff00016_Aln_CM020039.1.g224.t1, Scaff00019_Aln_CM020042.1.g116.t1, Scaff00023_Aln_CM020046.1.g217.t1, Scaff00025_Aln_CM020048.1.g354.t1, Scaff00031_Aln_CM020054.1.g9.t1, Scaff00040_Aln_CM020063.1.g405.t1, Scaff00040_Aln_CM020063.1.g456.t1, Scaff00040_Aln_CM020063.1.g514.t1, Scaff00040_Aln_CM020063.1.g747.t1, Scaff00040_Aln_CM020063.1.g926.t1, Scaff00040_Aln_CM020063.1.g1372.t1, Scaff00040_Aln_CM020063.1.g1484.t1, Scaff00040_Aln_CM020063.1.g1501.t1, Scaff00040_Aln_CM020063.1.g1931.t1, Scaff00040_Aln_CM020063.1.g1998.t1, Scaff00040_Aln_CM020063.1.g2123.t1, Scaff00041_Aln_CM020064.1.g906.t1, Scaff00041_Aln_CM020064.1.g932.t1		
		CM020024.1.g178.t1, CM020024.1.g353.t1, CM020024.1.g786.t1, CM020024.1.g829.t1, CM020024.1.g924.t1, CM020024.1.g954.t1, CM020024.1.g1130.t1, CM020024.1.g1209.t1, CM020024.1.g1648.t1, CM020024.1.g1747.t1,		
		CM020024.1.g1900.t1, CM020024.1.g2148.t1, CM020024.1.g2167.t1, CM020024.1.g2213.t1, CM020024.1.g2334.t1, CM020024.1.g2448.t1, CM020024.1.g2702.t1, CM020024.1.g2713.t1, CM020024.1.g2730.t1, CM020024.1.g3041.t1, CM020024.1.g3153.t1, CM020025.1.g20.t1, CM020025.1.g483.t1, CM020025.1.g730.t1, CM020025.1.g843.t1, CM020025.1.g1167.t1, CM020025.1.g1209.t1, CM020025.1.g2131.t1, CM020026.1.g61.t1, CM020026.1.g556.t1, CM020026.1.g718.t1, CM020026.1.g1089.t1, CM020026.1.g1108.t1, CM020026.1.g1465.t1, CM020026.1.g1710.t1, CM020027.1.g18.t1, CM020027.1.g434.t1, CM020027.1.g668.t1, CM020027.1.g730.t1, CM020027.1.g987.t1, CM020028.1.g51.t1, CM020030.1.g324.t1, CM020031.1.g338.t1, CM020031.1.g415.t1, CM020032.1.g655.t1, CM020033.1.g24.t1, CM020033.1.g126.t1, CM020033.1.g158.t1, CM020035.1.g370.t1, CM020036.1.g404.t1, CM020036.1.g440.t1, CM020036.1.g472.t1, CM020039.1.g237.t1, CM020042.1.g124.t1, CM020042.1.g425.t1, CM020046.1.g223.t1, CM020048.1.g358.t1, CM020049.1.g4.t1, CM020054.1.g12.t1, CM020061.1.g115.t1, CM020064.1.g1224.t1, CM020064.1.g1361.t1, CM020064.1.g1487.t1, CM020063.1.g503.t1, CM020063.1.g566.t1, CM020063.1.g924.t1,		

		CM020063.1.g1157.t1, CM020063.1.g1274.t1, CM020063.1.g1754.t1, CM020063.1.g1908.t1, CM020063.1.g2174.t1, CM020063.1.g2491.t1, CM020063.1.g2586.t1, CM020063.1.g2730.t1		
		NW_010387828.1.g1.t1, NW_010388345.1.g1.t1, NW_010391075.1.g1.t1, NW_010391380.1.g1.t1,		
		NW_010395901.1.g4.t1, NW_010396127.1.g1.t1, NW_010397189.1.g1.t1, NW_010397529.1.g1.t1, NW_010399354.1.g1.t1, NW_010399428.1.g3.t1, NW_010400434.1.g1.t1, NW_010408685.1.g1.t1, NW_010408737.1.g1.t1, NW_010411006.1.g1.t1, NW_010411028.1.g2.t1, NW_010413240.1.g2.t1, NW_010413479.1.g1.t1, NW_010414696.1.g1.t1, NW_010417447.1.g1.t1, NW_010422734.1.g1.t1, NW_010423775.1.g3.t1, NW_010423868.1.g3.t1, NW_010425247.1.g1.t1, NW_010430843.1.g3.t1, NW_010432016.1.g1.t1		
		JAODLK010000001.1.g830.t1, JAODLK010000001.1.g1314.t1, JAODLK010000001.1.g1339.t1, JAODLK010000001.1.g1391.t1, JAODLK010000001.1.g1577.t1, JAODLK010000023.1.g39.t1, JAODLK010000023.1.g262.t1, JAODLK010000024.1.g588.t1, JAODLK010000024.1.g1266.t1, JAODLK010000035.1.g771.t1, JAODLK010000035.1.g784.t1, JAODLK010000035.1.g1043.t1, JAODLK010000035.1.g1220.t1, JAODLK010000037.1.g66.t1, JAODLK010000046.1.g319.t1, JAODLK010000046.1.g362.t1, JAODLK010000046.1.g457.t1, JAODLK010000046.1.g619.t1, JAODLK010000048.1.g462.t1, JAODLK010000048.1.g490.t1, JAODLK010000048.1.g925.t1, JAODLK010000048.1.g1268.t1, JAODLK010000048.1.g1336.t1, JAODLK010000048.1.g1493.t1, JAODLK010000057.1.g338.t1, JAODLK010000057.1.g356.t1, JAODLK010000057.1.g471.t1, JAODLK010000057.1.g483.t1, JAODLK010000076.1.g67.t1, JAODLK010000557.1.g1.t1, JAODLK010001074.1.g1.t1		
	Diacylglycerol kinase	Scaff00001_Aln_CM020024.1.g2386.t1 CM020024.1.g2434.t1	T	K00901

	Diacylglycerol kinase beta	Scaff00005_Aln_CM020028.1.g759.t1 CM020028.1.g759.t1	T	K00901
	Diacylglycerol O-acyltransferase 1	Scaff00010_Aln_CM020033.1.g749.t1 CM020033.1.g784.t1	I	K11155
	Membrane-bound O-acyltransferase domain-containing protein	Scaff00001_Aln_CM020024.1.g925.t1 CM020024.1.g957.t1	I	K13517
Glycerophospholipid metabolism	Diacylglycerol kinase	Scaff00001_Aln_CM020024.1.g2386.t1 CM020024.1.g2434.t1	T	K00901
	Diacylglycerol kinase beta	Scaff00005_Aln_CM020028.1.g759.t1 CM020028.1.g759.t1	T	K00901
	Phosphate acyltransferases	Scaff00055_Aln_WNMG01000056.1.g5.t1 WNMG01000056.1.g3.t1	I	K13511
	Membrane-bound O-acyltransferase domain-containing protein	Scaff00001_Aln_CM020024.1.g925.t1 CM020024.1.g957.t1	I	K13517
	Phospholipase B1	Scaff00022_Aln_CM020045.1.g18.t1 CM020045.1.g28.t1	I	K14621
	Phospholipase A2	CM020052.1.g166.t1	ITU	K16342
Ether lipid metabolism	Phospholipase B1	Scaff00022_Aln_CM020045.1.g18.t1 CM020045.1.g28.t1	I	K14621
	Phospholipase A2	CM020052.1.g166.t1	ITU	K16342
Sphingolipid metabolism	Ceramide synthase 6	Scaff00006_Aln_CM020029.1.g361.t1 CM020029.1.g387.t1	U	K04710
Arachidonic acid metabolism	15-hydroxyprostaglandin dehydrogenase [NAD(+)]	Scaff00001_Aln_CM020024.1.g1605.t1, Scaff00002_Aln_CM020025.1.g1795.t1, Scaff00010_Aln_CM020033.1.g73.t1, Scaff00010_Aln_CM020033.1.g208.t1, Scaff00010_Aln_CM020033.1.g310.t1, Scaff00010_Aln_CM020033.1.g412.t1, Scaff00019_Aln_CM020042.1.g183.t1, Scaff00041_Aln_CM020064.1.g97.t1, Scaff00017_Aln_CM020040.1.g63.t1 CM020024.1.g1635.t1, CM020025.1.g1874.t1, CM020033.1.g85.t1, CM020033.1.g221.t1, CM020033.1.g335.t1, CM020033.1.g442.t1, CM020064.1.g131.t1, CM020040.1.g55.t1 NW_010388016.1.g2.t1, NW_010390261.1.g1.t1, NW_010390376.1.g2.t1, NW_010394842.1.g1.t1, NW_010395160.1.g2.t1, NW_010395732.1.g1.t1, NW_010405264.1.g1.t1, NW_010407427.1.g2.t1, NW_010412310.1.g1.t1, NW_010431939.1.g1.t1 JAODLK010000002.1.g632.t1, JAODLK010000024.1.g390.t1, JAODLK010000024.1.g746.t1, JAODLK010000035.1.g512.t1, JAODLK010000037.1.g8.t1, JAODLK010000048.1.g287.t1, JAODLK010000048.1.g1379.t1,	Q	K00069

		JAODLK010000066.1.g313.t1, JAODLK010000076.1.g703.t1		
	Phospholipase B1	Scaff00022_Aln_CM020045.1.g18.t1		
		CM020045.1.g28.t1	I	K14621
	Phospholipase A2	CM020052.1.g166.t1	ITU	K16342
	hematopoietic prostaglandin D synthase	NW_010395806.1.g3.t1		
		JAODLK010000035.1.g1058.t1	O	K04097
Linoleic acid metabolism and alpha-Linolenic acid metabolism		Scaff00022_Aln_CM020045.1.g18.t1		
	Phospholipase B1	CM020045.1.g28.t1	I	K14621
	Phospholipase A2	CM020052.1.g166.t1	ITU	K16342
Biosynthesis of unsaturated fatty acids	Very long chain fatty acid elongase 6	Scaff00003_Aln_CM020026.1.g1598.t1		
		CM020026.1.g1637.t1	I	K10203
	Peroxisomal multifunctional enzyme type 2	Scaff00040_Aln_CM020063.1.g1395.t1		
		CM020063.1.g1794.t1	Q	K12405

Appendix Table 9: Shared proteins in the accessory genome involved in the excretory system of the hornbill species. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Pathway	Protein name	Locus tags	COG category	KEGG KO term
Vasopressin-regulated water reabsorption	Cyclic AMP-responsive element-binding protein 3-like protein 4	Scaff00019_Aln_CM020042.1.g248.t1 CM020061.1.g111.t1	K	K09048
	Dynactin subunit	Scaff00027_Aln_CM020050.1.g576.t1 CM020050.1.g611.t1	Z	K10424
Aldosterone-regulated sodium reabsorption	Corticosteroid 11-beta-dehydrogenase isozyme 2	Scaff00017_Aln_CM020040.1.g63.t1 CM020040.1.g55.t1	Q	K00071
Endocrine and other factor-regulated calcium reabsorption	AP-2 complex subunit alpha-2	Scaff00020_Aln_CM020043.1.g468.t1	U	K11824
	AP-2 complex subunit sigma	Scaff00071_Aln_WNMG01000080.1.g20.t1 WNMG01000080.1.g27.t1	U	K11827
	calretinin	Scaff00017_Aln_CM020040.1.g53.t1 CM020040.1.g45.t1	T	K14757
	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase	JAODLK010000081.1.g92.t1	I	K05858
	Dynamin-3	NW_010417008.1.g2.t1	U	K01528
Proximal tubule bicarbonate reclamation	Glutamate dehydrogenase 1, mitochondrial-like	Scaff00018_Aln_CM020041.1.g219.t1 CM020041.1.g217.t1	E	K00261
Collecting duct acid secretion	proton-transporting ATPase activity, rotational mechanism	Scaff00002_Aln_CM020025.1.g55.t1 CM020025.1.g57.t1	C	K02151
	V-type proton ATPase subunit e	Scaff00009_Aln_CM020032.1.g408.t1 CM020032.1.g388.t1	C	K02153

Appendix Table 10: Shared proteins in the accessory genome involved in the immune system of the hornbill species. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Pathway	Protein name	Locus tags	COG category	KEGG KO term
NOD-like receptor signalling pathway	inositol 1,4,5-trisphosphate receptor type 1	Scaff00014_Aln_CM020037.1.g475.t1 CM020037.1.g487.t1	T	K04958
	interferon alpha/beta receptor 1	Scaff00007_Aln_CM020030.1.g486.t1 CM020030.1.g481.t1	T	K05130
	Inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	Scaff00039_Aln_CM020062.1.g100.t1 CM020062.1.g129.t1	K	K07210
	Cytochrome b-245 light	Scaff00017_Aln_CM020040.1.g371.t1 CM020040.1.g357.t1	C	K08009
	interleukin-8	Scaff00023_Aln_CM020046.1.g21.t1 CM020046.1.g25.t1	T	K10030
	C-C motif	Scaff00027_Aln_CM020050.1.g8.t1 CM020050.1.g8.t1	G	K12499
	NLR family CARD domain-containing protein 4	CM020031.1.g814.t1	O	K12805
	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase	JAODLK010000081.1.g92.t1 NW_010392462.1.g2.t1	I	K05858
IL-17 signalling pathway B cell receptor signaling pathway	transcription factor jun-D	Scaff00032_Aln_CM020055.1.g448.t1 CM020055.1.g445.t1	K	K04449
	Inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	Scaff00039_Aln_CM020062.1.g100.t1 CM020062.1.g129.t1	K	K07210
	interleukin-8	Scaff00023_Aln_CM020046.1.g21.t1 CM020046.1.g25.t1	T	K10030
	Serine arginine-rich splicing factor 1	Scaff00027_Aln_CM020050.1.g269.t1 CM020050.1.g270.t1	A	K12890
	S-100/ICaBP type calcium binding domain	Scaff00038_Aln_CM020061.1.g265.t1	S	K21128
	Interleukin-17	JAODLK010000001.1.g2200.t1 NW_010387957.1.g2.t1	S	K05492
B cell receptor signaling pathway	RAC-beta serine/threonine-protein kinase isoform X1	Scaff00084_Aln_WNMG01000093.1.g16.t1 WNMG01000093.1.g21.t1	T	K04456
	Immunoglobulin heavy constant epsilon	Scaff00069_Aln_WNMG01000078.1.g6.t1 WNMG01000078.1.g9.t1	S	K06856
	Ig heavy chain Mem5-like	Scaff00069_Aln_WNMG01000078.1.g25.t1, Scaff00069_Aln_WNMG01000078.1.g26.t1, Scaff00069_Aln_WNMG01000078.1.g30.t1, Scaff00069_Aln_WNMG01000078.1.g31.t1 WNMG01000078.1.g24.t1, WNMG01000078.1.g25.t1, WNMG01000078.1.g29.t1, WNMG01000078.1.g30.t1 JAODLK010001779.1.g1.t1, JAODLK010007728.1.g1.t1 NW_010399493.1.g1.t1, NW_010423276.1.g1.t1	S	K06856
		Scaff00039_Aln_CM020062.1.g100.t1		

	Inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	CM020062.1.g129.t1	K	K07210
	B-cell lymphoma leukemia 10	Scaff00011_Aln_CM020034.1.g421.t1		
		CM020034.1.g420.t1	S	K07368
		Scaff00009_Aln_CM020032.1.g596.t1		
		CM020032.1.g573.t1		
		Scaff00021_Aln_CM020044.1.g498.t1		
		CM020044.1.g494.t1		
		Scaff00027_Aln_CM020050.1.g396.t1		
		CM020050.1.g395.t1		
		Scaff00019_Aln_CM020042.1.g294.t1, Scaff00034_Aln_CM020057.1.g1.t1, Scaff00037_Aln_CM020060.1.g14.t1, Scaff00038_Aln_CM020061.1.g96.t1, Scaff00095_Aln_WNMG01000104.1.g1.t1, Scaff00102_Aln_WNMG01000112.1.g1.t1, scaffold5904_size31777.g4.t1, scaffold6352_size26915.g1.t1, scaffold6393_size26530.g4.t1, scaffold7208_size19954.g2.t1, scaffold7498_size18197.g1.t1, scaffold7794_size16711.g1.t1, scaffold7794_size16711.g2.t1, scaffold7966_size15989.g1.t1, scaffold8027_size15721.g1.t1, scaffold8057_size15556.g1.t1, scaffold8108_size15311.g1.t1, scaffold8725_size13083.g1.t1, scaffold8728_size13076.g1.t1, scaffold8795_size12853.g1.t1, scaffold8863_size12608.g1.t1, scaffold9191_size11573.g1.t1, scaffold9515_size10532.g1.t1, scaffold9649_size10141.g1.t1, scaffold9684_size10053.g1.t1, scaffold9723_size9962.g1.t1, scaffold9752_size9871.g1.t1		
		CM020042.1.g299.t1, CM020057.1.g1.t1, CM020059.1.g1.t1, CM020060.1.g8.t1, CM020061.1.g112.t1, CM020061.1.g334.t1, WNMG01000099.1.g1.t1, WNMG01000104.1.g1.t1		
		Scaff00039_Aln_CM020062.1.g100.t1		
		CM020062.1.g129.t1		
		Scaff00007_Aln_CM020030.1.g514.t1		
		CM020030.1.g505.t1		
	Phosphoinositide 3-kinase adapter protein	JAODLK010000048.1.g1068.t1, JAODLK010002843.1.g1.t1		
		NW_010424057.1.g1.t1	T	K12230

Appendix Table 11: Shared proteins in the accessory genome involved in the endocrine and digestive system of the hornbill species. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Pathway	Protein name	Locus tags	COG category	KEGG KO term
Insulin secretion	Cyclic AMP-responsive element-binding protein 3-like protein 4	Scaff00019_Aln_CM020042.1.g248.t1 CM020061.1.g111.t1	K	K09048
	Adenylate cyclase type 2	JAODLK010000046.1.g403.t1 NW_010414497.1.g3.t1	C	K08042
	1-phosphatidylinositol 4,5-bisphosphate phosphodiesterase	JAODLK010000081.1.g92.t1 NW_010392462.1.g2.t1	I	K05858
	voltage-dependent L-type calcium channel subunit alpha-1D	JAODLK010000018.1.g78.t1 NW_010423251.1.g2.t1	P	K04851
	Piccolo Zn-finger	JAODLK010000024.1.g228.t1 NW_010404191.1.g1.t1	U	K16882
	Cyclic AMP-dependent transcription factor ATF-4	JAODLK010000024.1.g1040.t1 NW_010410838.1.g1.t1	K	K04374
	Potassium large conductance calcium-activated channel, subfamily M, alpha member	NW_010427549.1.g2.t1	P	K04936
Protein digestion and absorption	Proton-coupled amino acid transporter 4	Scaff00001_Aln_CM020024.1.g3004.t1 CM020024.1.g3076.t1	E	K14209
	Potassium voltage-gated channel subfamily KQT member 1-like	Scaff00002_Aln_CM020025.1.g1984.t1 CM020025.1.g2068.t1	P	K04926
	Collagen alpha-5(IV)	Scaff00016_Aln_CM020039.1.g121.t1	W	K06237
	Collagen alpha-3(IX) chain	NW_010400860.1.g2.t1	W	K08131

Appendix Table 12: Shared proteins in the accessory genome involved in the environmental adaptation and development and regeneration of the hornbill species. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Pathway	Protein name	Locus tags	COG category	KEGG KO term
	Cyclic AMP-responsive element-binding protein 3-like protein 4	Scaff00019_Aln_CM020042.1.g248.t1 CM020061.1.g111.t1	K	K09048
	Ribosomal protein S6 kinase	Scaff00007_Aln_CM020030.1.g765.t1 CM020030.1.g753.t1	T	K04373
	ribosomal protein S6 kinase activity	Scaff00020_Aln_CM020043.1.g159.t1 CM020043.1.g158.t1	T	K04688
	NADH dehydrogenase ubiquinone 1 subunit C2	Scaff00001_Aln_CM020024.1.g2822.t1 CM020024.1.g2897.t1	C	K03968
	NADH dehydrogenase (ubiquinone) 1 beta subcomplex	Scaff00002_Aln_CM020025.1.g1628.t1 CM020025.1.g1721.t1	C	K03958
	cytochrome b-c1 complex subunit 8 isoform X1	Scaff00018_Aln_CM020041.1.g44.t1 CM020041.1.g41.t1	C	K00418
	NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 2	Scaff00018_Aln_CM020041.1.g488.t1 CM020041.1.g483.t1	C	K03946
	ATP synthase, H transporting, mitochondrial Fo complex, subunit C2 (subunit 9)	Scaff00027_Aln_CM020050.1.g496.t1 CM020050.1.g509.t1	C	K02128
	NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 13	Scaff00032_Aln_CM020055.1.g409.t1 CM020055.1.g406.t1	CD	K11353
	Mitochondrial ATP synthase g subunit	Scaff00035_Aln_CM020058.1.g232.t1 CM020058.1.g224.t1	C	K02140
	Cytochrome c oxidase assembly factor 4 homolog	Scaff00001_Aln_CM020024.1.g2733.t1 CM020024.1.g2808.t1	S	K18177
	cytochrome c oxidase assembly	Scaff00026_Aln_CM020049.1.g297.t1 CM020049.1.g307.t1	S	K18175
	regulation of glucose mediated signaling pathway	Scaff00041_Aln_CM020064.1.g115.t1	K	K11647
	Adenylate cyclase type 2	JAODLK010000046.1.g403.t1 NW_010414497.1.g3.t1	C	K08042
	NADH dehydrogenase ubiquinone 1 alpha subcomplex subunit 10, mitochondrial	JAODLK010000066.1.g140.t1 NW_010432142.1.g1.t1	C	K03954
Thermogenesis	cytochrome c oxidase assembly	JAODLK010000026.1.g250.t1	C	K18183
	RAC-beta serine/threonine-protein kinase isoform X1	Scaff00084_Aln_WNMG01000093.1.g16.t1 WNMG01000093.1.g21.t1	T	K04456
	Inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	Scaff00039_Aln_CM020062.1.g100.t1 CM020062.1.g129.t1	K	K07210
	inositol 1,4,5-trisphosphate receptor type 1	Scaff00014_Aln_CM020037.1.g475.t1 CM020037.1.g487.t1	T	K04958
Osteoclast differentiation	Interferon alpha beta	Scaff00007_Aln_CM020030.1.g486.t1 CM020030.1.g481.t1	T	K05130

	Cytochrome b-245 light	Scaff00017_Aln_CM020040.1.g371.t1 CM020040.1.g357.t1	C	K08009
	Calcium calmodulin-dependent protein kinase	Scaff00040_Aln_CM020063.1.g1532.t1 CM020063.1.g1958.t1	T	K05869
	Transcription factor RelB	Scaff00090_Aln_WNMG01000099.1.g2.t1 WNMG01000099.1.g6.t1	S	K09253

Appendix Table 13: Shared proteins in the accessory genome involved in environmental information processing pathways. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Pathway	Protein name	Locus tags	COG category	KEGG KO term
	Ras GTPase-activating protein	Scaff00040_Aln_CM020063.1.g1115.t1, Scaff00041_Aln_CM020064.1.g135.t1, Scaff00041_Aln_CM020064.1.g137.t1, Scaff00041_Aln_CM020064.1.g142.t1 CM020064.1.g174.t1, CM020064.1.g176.t1, CM020064.1.g177.t1, CM020064.1.g189.t1, CM020063.1.g1395.t1	T	K04352
	fibroblast growth factor 10	Scaff00040_Aln_CM020063.1.g412.t1 CM020063.1.g512.t1	T	K04358
	RAC-beta serine/threonine-protein kinase isoform X1	Scaff00084_Aln_WNMG01000093.1.g16.t1 WNMG01000093.1.g21.t1	T	K04456
	Inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma	Scaff00039_Aln_CM020062.1.g100.t1 CM020062.1.g129.t1	K	K07210
Ras signalling pathway	Ras-related protein R-Ras2	Scaff00001_Aln_CM020024.1.g488.t1, Scaff00001_Aln_CM020024.1.g634.t1, Scaff00001_Aln_CM020024.1.g1287.t1, Scaff00001_Aln_CM020024.1.g3490.t1, Scaff00002_Aln_CM020025.1.g213.t1, Scaff00005_Aln_CM020028.1.g676.t1, Scaff00005_Aln_CM020028.1.g699.t1, Scaff00007_Aln_CM020030.1.g908.t1, Scaff00009_Aln_CM020032.1.g671.t1, Scaff00010_Aln_CM020033.1.g399.t1, Scaff00041_Aln_CM020064.1.g521.t1, Scaff00041_Aln_CM020064.1.g847.t1, Scaff00041_Aln_CM020064.1.g935.t1, Scaff00068_Aln_WNMG01000077.1.g37.t1 CM020024.1.g498.t1, CM020024.1.g1310.t1, CM020024.1.g3547.t1, CM020024.1.g3915.t1, CM020025.1.g231.t1, CM020026.1.g687.t1, CM020028.1.g701.t1, CM020029.1.g636.t1, CM020030.1.g879.t1, CM020031.1.g888.t1, CM020032.1.g653.t1, CM020033.1.g427.t1, CM020064.1.g149.t1, CM020064.1.g789.t1, WNMG01000077.1.g32.t1 JAODLK010000001.1.g404.t1, JAODLK010000001.1.g2223.t1, JAODLK010000013.1.g365.t1, JAODLK010000023.1.g134.t1, JAODLK010000024.1.g493.t1, JAODLK010000035.1.g976.t1, JAODLK010000035.1.g1198.t1, JAODLK010000048.1.g407.t1, JAODLK010000048.1.g587.t1, JAODLK010000048.1.g1348.t1, JAODLK010000048.1.g1473.t1, JAODLK010000057.1.g590.t1, JAODLK010000076.1.g164.t1	S	K07830

		NW_010389080.1.g3.t1, NW_010389992.1.g1.t1, NW_010397045.1.g1.t1, NW_010397171.1.g1.t1, NW_010398498.1.g1.t1, NW_010404100.1.g2.t1, NW_010407605.1.g3.t1, NW_010416713.1.g1.t1, NW_010416899.1.g2.t1, NW_010417166.1.g1.t1, NW_010425158.1.g1.t1, NW_010428480.1.g4.t1, NW_010428788.1.g1.t1		
	Forkhead box protein O4	Scaff00016_Aln_CM020039.1.g340.t1		
		CM020039.1.g347.t1	K	K12358
	Insulin-like growth factor	Scaff00020_Aln_CM020043.1.g250.t1		
		CM020043.1.g246.t1	T	K13769
	Phospholipase A2	CM020052.1.g166.t1	ITU	K16342
	guanine nucleotide-binding protein G(I)/G(S)/G(O) subunit gamma-10	CM020063.1.g1712.t1	T	K04545
	growth factor D	Scaff00001_Aln_CM020024.1.g3161.t1		K05450
		CM020024.1.g3234.t1		
		NW_010422557.1.g1.t1	T	
JAK-STAT signalling pathway	RAC-beta serine/threonine-protein kinase isoform X1	Scaff00084_Aln_WNMG01000093.1.g16.t1		
		WNMG01000093.1.g21.t1	T	K04456
	Histone acetyltransferase	Scaff00002_Aln_CM020025.1.g1394.t1		
		CM020025.1.g1482.t1	K	K04498
	Granulocyte-macrophage colony-stimulating factor receptor subunit alpha-like	Scaff00005_Aln_CM020028.1.g639.t1		
		CM020042.1.g483.t1	T	K05066
	interleukin-21 receptor	Scaff00021_Aln_CM020044.1.g498.t1		
		CM020044.1.g494.t1	T	K05075
	Cytokine receptor-like factor 2	Scaff00001_Aln_CM020024.1.g1647.t1		
		CM020024.1.g1682.t1	T	K05078
	interferon alpha/beta receptor 1	Scaff00007_Aln_CM020030.1.g486.t1		
		CM020030.1.g481.t1	T	K05130
	Leukemia inhibitory factor	Scaff00025_Aln_CM020048.1.g389.t1		
		CM020048.1.g394.t1	T	K05419
	Interleukin-23 subunit alpha	Scaff00027_Aln_CM020050.1.g396.t1		
		CM020050.1.g395.t1	T	K05426
AMPK signalling pathway	fatty acid synthase	Scaff00028_Aln_CM020051.1.g9.t1		
		CM020051.1.g9.t1	Q	K00665
	glycogen [starch] synthase, liver isoform X1	Scaff00002_Aln_CM020025.1.g1868.t1		
		CM020025.1.g1947.t1	G	K00693
	RAC-beta serine/threonine-protein kinase isoform X1	Scaff00084_Aln_WNMG01000093.1.g16.t1		
		WNMG01000093.1.g21.t1	T	K04456
	ribosomal protein S6 kinase activity	Scaff00020_Aln_CM020043.1.g159.t1		
		CM020043.1.g158.t1	T	K04688
	Cyclic AMP-responsive element-binding protein 3-like protein 4	Scaff00019_Aln_CM020042.1.g248.t1		
		CM020061.1.g111.t1	K	K09048
	Eukaryotic translation elongation factor 2	JAODLK010000048.1.g75.t1		
		NW_010407400.1.g1.t1	J	K03234

Appendix Table 14: Shared proteins in the accessory genome involved in cellular processes pathways. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Protein name	Locus tags	COG category	KEGG KO term
Ribosomal protein S6 kinase	Scaff00007_Aln_CM020030.1.g765.t1 CM020030.1.g753.t1	T	K04373
inositol 1,4,5-trisphosphate receptor type 1	Scaff00014_Aln_CM020037.1.g475.t1 CM020037.1.g487.t1	T	K04958
14-3-3 protein gamma	Scaff00027_Aln_CM020050.1.g136.t1 CM020050.1.g142.t1	O	K16198
anaphase-promoting complex subunit 11	Scaff00028_Aln_CM020051.1.g23.t1 CM020051.1.g23.t1	O	K03358
shugoshin 1 isoform X1	Scaff00031_Aln_CM020054.1.g158.t1 CM020054.1.g177.t1	K	K11580
Adenylate cyclase type 2	JAODLK010000046.1.g403.t1 NW_010414497.1.g3.t1	C	K08042

Appendix Table 15: Paralogous proteins involved in cellular community pathways. Protein locus tags in blue represent those identified in the *Bucorvus* species and those in orange in the *Buceros* species.

Pathway	Protein name	Locus tag	COG category	KEGG KO term
	tenascin-R isoform X1	Scaff00011_Aln_CM020034.1.g30.t1	T	K06252
	Laminin subunit alpha-3	Scaff00029_Aln_CM020052.1.g226.t1	W	K06240
Focal adhesion	Phosphatidylinositol 3-kinase regulatory subunit	CM020063.1.g787.t1	T	K02649
	MARVEL domain-containing protein 2-like, partial	CM020063.1.g1225.t1	K	K17291
Tight junction	Protein kinase C, epsilon	NW_010418314.1.g1.t1	T	K18050
Signaling pathways regulating pluripotency of stem cells	Phosphatidylinositol 3-kinase regulatory subunit	CM020063.1.g787.t1	T	K02649

Appendix Table 16: COG classifications of the core and unique SGH and AGH orthogroups and those with paralogous proteins

COG category	Function	Core	Southern Ground Hornbill unique	Abyssinian Ground Hornbill unique	Southern Ground Hornbill paralogs	Abyssinian Ground Hornbill paralogs
Information storage and processing						
A	RNA processing and modification	488	42	37	0	3
B	Chromatin Structure and dynamics	225	26	16	0	3
J	Translation	380	29	33	12	4
K	Transcription	1,657	253	116	32	27
L	Replication and repair	1,061	456	108	122	458
Cellular processes and signalling						
D	Cell cycle control and mitosis	303	45	17	0	4
M	Cell wall/membrane/envelop biogenesis	87	12	3	0	0
N	Cell motility	17	4	1	0	0
O	Post-translational modification, protein turnover, chaperone functions	1,613	241	115	127	168
T	Signal Transduction	3,931	459	216	116	52
U	Intracellular trafficking and secretion	761	70	53	1	3
V	Defence mechanism	201	50	12	27	2
W	Extracellular structures	276	33	17	0	2
X	Mobilome	0	0	0	0	0
Y	Nuclear structure	43	6	3	0	2
Z	Cytoskeleton	639	77	32	12	10
Metabolism						
C	Energy production and conversion	339	41	16	89	5
E	Amino Acid metabolism and transport	2,117	473	148	82	245
F	Nucleotide metabolism and transport	194	12	5	13	7
G	Carbohydrate metabolism and transport	636	110	39	10	24
H	Coenzyme metabolism	90	6	8	0	0
I	Lipid metabolism	441	43	18	2	3
P	Inorganic ion transport and metabolism	520	52	36	0	0
Q	Secondary Structure	229	24	6	11	3
Poorly characterized						
R	General function prediction only	0	0	0	0	0
S	Function Unknown	10,456	2,894	693	584	2,225
-	Undetermined	142	63	18	7	13
Total		**26,839**	**5,521**	**1,766**	**1,247**	**3,263**

Appendix Table 17: KEGG pathways reconstructed for the unique SGH and AGH orthogroups

KEGG class	Category	Southern Ground Hornbill	Abyssinian Ground Hornbill	Factor
Metabolism	Global and overview maps	159	90	1,766667
	Carbohydrate metabolism	39	12	3,25
	Energy metabolism	13	10	1,3
	Lipid metabolism	33	9	3,666667
	Nucleotide metabolism	12	5	2,4
	Amino acid metabolism	21	14	1,5
	Metabolism of other amino acids	9	3	3
	Glycan biosynthesis and metabolism	23	11	2,090909
	Metabolism of cofactors and vitamins	11	10	1,1
	Metabolism of terpenoids and polyketides	1	2	0,5
	Biosynthesis of other secondary metabolites	6	4	1,5
	Xenobiotics biodegradation and metabolism	9	3	3
Genetic Information Processing	Transcription	9	7	1,285714
	Translation	28	16	1,75
	Folding, sorting and degradation	28	27	1,037037
	Replication and repair	8	21	0,380952
	Chromosome	14	6	2,333333
	Information processing in viruses	2	8	0,25
Environmental Information Processing	Membrane transport	0	1	0
	Signal transduction	243	172	1,412791
	Signalling molecules and interaction	35	20	1,75
Cellular Processes	Transport and catabolism	66	43	1,534884
	Cell growth and death	59	53	1,113208
	Cellular community - eukaryotes	47	27	1,740741
	Cellular community - prokaryotes	1	1	1
	Cell motility	52	25	2,08
Organismal Systems	Immune system	127	102	1,245098
	Endocrine system	108	96	1,125
	Circulatory system	28	14	2
	Digestive system	39	22	1,772727
	Excretory system	10	8	1,25
	Nervous system	53	49	1,081633
	Sensory system	20	12	1,666667
	Development and regeneration	28	23	1,217391
	Aging	10	10	1
	Environmental adaptation	18	17	1,058824

Human Diseases	Cancer: overview	110	79	1,392405
	Cancer: specific types	85	49	1,734694
	Infectious disease: viral	118	86	1,372093
	Infectious disease: bacterial	86	73	1,178082
	Infectious disease: parasitic	34	30	1,133333
	Immune disease	34	27	1,259259
	Neurodegenerative disease	137	124	1,104839
	Substance dependence	24	20	1,2
	Cardiovascular disease	62	43	1,44186
	Endocrine and metabolic disease	43	39	1,102564
	Drug resistance: antineoplastic	14	12	1,166667
Total		2116	1535	

Appendix Table 18: SGH and AGH species-specific proteins functioning within pathways functioning in metabolism. SGH proteins are represented in pink and the AGH in cyan.

Category	Pathways	Predicted protein	Locus tag	COG category	KEGG KO term
Carbohydrate metabolism	Amino sugar and nucleotide sugar metabolism	Glucosamine-6-phosphate	Scaff00003_Aln_CM020026.1.g786.t1	G	K02564
		Belongs to the hexokinase family	Scaff00009_Aln_CM020032.1.g894.t1	G	K00844
		GDP-L-fucose synthase	Scaff00010_Aln_CM020033.1.g777.t1	GO	K02377
		Fucose-1-phosphate guanylyltransferase	Scaff00011_Aln_CM020034.1.g774.t1	S	K00976
		L-fucokinase	scaffold2103_size170777.g3.t1	G	K05305
		Bifunctional UDP-N-acetylglucosamine 2-epimerase N-acetylmannosamine kinase	CM020063.1.g2220.t1	GK	K12409
Lipid metabolism	Glycerolipid metabolism	1-acyl-sn-glycerol-3-phosphate acyltransferase epsilon	Scaff00001_Aln_CM020024.1.g706.t1	I	K19007
		Belongs to the FGGY kinase family	Scaff00001_Aln_CM020024.1.g2037.t1, Scaff00001_Aln_CM020024.1.g2460.t1, Scaff00001_Aln_CM020024.1.g852.t1, Scaff00003_Aln_CM020026.1.g1220.t1, Scaff00003_Aln_CM020026.1.g1615.t1, Scaff00003_Aln_CM020026.1.g1719.t1, Scaff00003_Aln_CM020026.1.g336.t1, Scaff00005_Aln_CM020028.1.g428.t1, Scaff00005_Aln_CM020028.1.g572.t1, Scaff00005_Aln_CM020028.1.g660.t1, Scaff00005_Aln_CM020028.1.g717.t1, Scaff00005_Aln_CM020028.1.g782.t1, Scaff00007_Aln_CM020030.1.g1021.t1, Scaff00007_Aln_CM020030.1.g196.t1, Scaff00007_Aln_CM020030.1.g220.t1, Scaff00007_Aln_CM020030.1.g420.t1, Scaff00008_Aln_CM020031.1.g599.t1, Scaff00010_Aln_CM020033.1.g107.t1, Scaff00010_Aln_CM020033.1.g114.t1, Scaff00010_Aln_CM020033.1.g140.t1, Scaff00016_Aln_CM020039.1.g214.t1, Scaff00022_Aln_CM020045.1.g288.t1, Scaff00022_Aln_CM020045.1.g295.t1, Scaff00032_Aln_CM020055.1.g635.t1, Scaff00040_Aln_CM020063.1.g1563.t1, Scaff00040_Aln_CM020063.1.g2011.t1, Scaff00040_Aln_CM020063.1.g778.t1, Scaff00041_Aln_CM020064.1.g101.t1, Scaff00041_Aln_CM020064.1.g11.t1, Scaff00041_Aln_CM020064.1.g524.t1, Scaff00041_Aln_CM020064.1.g57.t1, Scaff00041_Aln_CM020064.1.g930.t1	G	K00864
		Diacylglycerol kinase iota	Scaff00002_Aln_CM020025.1.g1644.t1	T	K00901
		Glycerol-3-phosphate acyltransferase	Scaff00023_Aln_CM020046.1.g272.t1	I	K13506
		Glycerol-3-phosphate acyltransferase 2	Scaff00036_Aln_CM020059.1.g68.t1	I	K00629

		Pancreatic lipase-related protein	Scaff00009_Aln_CM020032.1.g176.t1, Scaff00009_Aln_CM020032.1.g177.t1	T	K14076
		Phosphatidate phosphatase LPIN2	Scaff00004_Aln_CM020027.1.g1064.t1	IN	K15728
		Belongs to the FGGY kinase family	CM020024.1.g1641.t1, CM020024.1.g2535.t1, CM020024.1.g2929.t1, CM020025.1.g1237.t1, CM020026.1.g476.t1, CM020026.1.g562.t1, CM020026.1.g600.t1, CM020026.1.g1478.t1, CM020028.1.g1119.t1, CM020030.1.g306.t1, CM020030.1.g999.t1, CM020052.1.g236.t1, CM020064.1.g1010.t1	G	K00864
	Glycerophospholipid metabolism	1-acyl-sn-glycerol-3-phosphate acyltransferase epsilon	Scaff00001_Aln_CM020024.1.g706.t1	I	K19007
		CDP-alcohol phosphatidyltransferase	Scaff00029_Aln_CM020052.1.g129.t1	I	K00993
		Cytosolic phospholipase A2 epsilon-like	Scaff00019_Aln_CM020042.1.g400.t1	ITU	K16342
		Diacylglycerol kinase iota	Scaff00002_Aln_CM020025.1.g1644.t1	T	K00901
		Glycerol-3-phosphate acyltransferase	Scaff00023_Aln_CM020046.1.g272.t1	I	K13506
		Glycerol-3-phosphate acyltransferase 2	Scaff00036_Aln_CM020059.1.g68.t1	I	K00629
		layer formation in cerebral cortex	Unaln_scaffold7359_size19031.g1.t1, Unaln_scaffold7359_size19031.g2.t1	S	K13516
		Phosphatidate phosphatase LPIN2	Scaff00004_Aln_CM020027.1.g1064.t1	IN	K15728
		Phosphatidylserine synthase 1	Scaff00010_Aln_CM020033.1.g405.t1	I	K08729
		Acyl-protein thioesterase 2	CM020057.1.g120.t1	I	K06130
Nucleotide metabolism	Pyrimidine metabolism	Cytosolic 5'-nucleotidase	Scaff00001_Aln_CM020024.1.g1103.t1	S	K01081
		Deoxyuridine 5'-triphosphate nucleotidohydrolase	Scaff00004_Aln_CM020027.1.g190.t1, Scaff00041_Aln_CM020064.1.g333.t1, Scaff00041_Aln_CM020064.1.g801.t1	F	K01520
		Dihydropyrimidine dehydrogenase NADP	Scaff00011_Aln_CM020034.1.g349.t1	F	K00207
		ectonucleotide pyrophosphatase phosphodiesterase	Scaff00001_Aln_CM020024.1.g47.t1	S	K01513
		Provides the precursors necessary for DNA synthesis. Catalyzes the biosynthesis of deoxyribonucleotides from the corresponding ribonucleotides	Scaff00001_Aln_CM020024.1.g2742.t1	F	K10807
		Uridine 5'-monophosphate synthase	Scaff00006_Aln_CM020029.1.g663.t1	F	K13421
		Uridine-cytidine kinase 2	Scaff00011_Aln_CM020034.1.g101.t1	TZ	K00876

Appendix Table 19: SGH and AGH species-specific proteins functioning within pathways mediating immune response. SGH proteins are represented in pink and the AGH in cyan.

Category	Pathways	Predicted protein	Locus tag	COG category	KEGG KO term
Transport and catabolism	Efferocytosis	Cation transporting ATPase, C-terminus	Scaff00061_Aln_WNMG01000065.1.g1.t1	P	K05853
		Lysosomal acid lipase cholesteryl ester hydrolase-like	Scaff00009_Aln_CM020032.1.g514.t1	I	K01052
		Monocarboxylate transporter 1	Scaff00032_Aln_CM020055.1.g285.t1	G	K08179
		nuclear factor of activated T-cells, cytoplasmic	Scaff00004_Aln_CM020027.1.g14.t1	K	K04446
		positive regulation of synapse maturation	Scaff00036_Aln_CM020059.1.g52.t1	T	K04515
		Vacuolar protein	Scaff00015_Aln_CM020038.1.g510.t1	U	K20182
		Vacuolar protein sorting-associated protein 16 homolog	Scaff00003_Aln_CM020026.1.g264.t1	U	K20180
		Vacuolar protein sorting-associated protein 8	Scaff00012_Aln_CM020035.1.g556.t1	O	K20178
Immune system	Complement and coagulation cascades	Complement factor I	Scaff00003_Aln_CM020026.1.g1590.t1	O	K01333
		Tissue factor pathway inhibitor	CM020029.1.g855.t1	O	K03909
		complement factor H	CM020034.1.g248.t1	T	K04004
		C1 inhibitor	CM020043.1.g16.t1	V	K04001
		CD59 glycoprotein	CM020043.1.g208.t1	S	K04008
		Complement component	CM020063.1.g475.t1	W	K03995
		receptor 1	CM020063.1.g852.t1	T	K03914

Appendix Table 20: KEGG pathways reconstructed for the gene duplicated SGH and AGH orthogroups. The overrepresented categories (3-fold differences) in the SGH are highlighted in pink and those in the AGH in cyan.

KEGG class	Category	Southern Ground Hornbill	Abyssinian Ground Hornbill	Factor
	Global and overview maps	7	6	1,17
	Carbohydrate metabolism	3	0	3
	Lipid metabolism	1	1	1
	Nucleotide metabolism	1	1	1
	Amino acid metabolism	1	1	1
	Glycan biosynthesis and metabolism	1	3	0,33
	Metabolism of cofactors and vitamins	0	1	0
Metabolism	Xenobiotics biodegradation and metabolism	1	1	1
Genetic information processing	Translation	3	4	0,75
	Folding, sorting and degradation	4	3	1,33
	Chromosome	0	3	0
	Membrane transport	1	0	1
	Signal transduction	25	22	1,14
	Signalling molecules and interaction	4	4	1
	Transport and catabolism	6	7	0,86
Environmental information processing	Cell growth and death	7	6	1,17
	Cellular community - eukaryotes	3	2	1,5
	Cell motility	8	5	1,6
	Immune system	25	21	1,19
	Endocrine system	7	15	0,47
	Circulatory system	3	2	1,5
	Digestive system	0	1	0
	Excretory system	1	1	1
	Nervous system	5	6	0,83
	Sensory system	3	1	3
	Development and regeneration	1	1	1
	Aging	0	3	0
Organism system	Environmental adaptation	1	3	0,33
	Cancer: overview	15	14	1,07
	Cancer: specific types	3	3	1
	Infectious disease: viral	13	13	1
	Infectious disease: bacterial	9	10	0,9
	Infectious disease: parasitic	8	5	1,6
	Immune disease	20	15	1,33
	Neurodegenerative disease	10	19	0,53
	Substance dependence	4	8	0,5
	Cardiovascular disease	10	3	3,33
	Endocrine and metabolic disease	5	4	1,25
Human diseases	Drug resistance: antineoplastic	1	0	1
Total		**220**	**218**	

Appendix Table 21: Proportion of genomic variants across the various regions of the gene and the impact these have on the protein-coding sequence.

Type of impact	SVs	INDELs	SNPs	Impact rating *
Bidirectional gene fusion	0.38	0.00	0.00	No rating
Conservative inframe deletion	0.19	0.03	0.00	Moderate
Conservative inframe insertion	0.28	0.04	0.00	Moderate
Disruptive inframe deletion	0.13	0.03	0.00	Moderate
Disruptive inframe insertion	0.11	0.02	0.00	Moderate
Downstream gene variant	9.11	11.63	12.17	Modifier
Exon loss variant	0.78	0.00	0.00	No rating
Feature ablation	0.21	0.00	0.00	No rating
Feature elongation	0.00	0.00	0.00	High
Feature fusion	28.79	0.00	0.00	No rating
Frameshift variant	2.12	1.14	0.00	High
Gene fusion	2.69	0.00	0.00	No rating
Intergenic region	22.73	45.77	45.05	Modifier
Initiator codon variant	0.00	0.00	0.00	No rating
Intron variant	13.55	29.62	27.95	Modifier
Missense variant	0.16	0.00	1.36	Moderate
Splice acceptor variant	0.26	0.02	0.01	High
Splice donor variant	0.21	0.01	0.01	High
Splice region variant	0.55	0.21	0.16	Low
Start lost	0.14	0.00	0.01	High
Start retained variant	0.01	0.00	0.00	Low
Stop gained	0.46	0.01	0.04	High
Stop lost	0.13	0.00	0.01	High
Stop retained variant	0.00	0.00	0.00	Low
Synonymous variant	0.77	0.00	0.99	Low
Transcript ablation	6.56	0.00	0.00	High
Upstream gene variant	9.72	11.45	12.25	Modifier

***** Impact rating as determined by Cingolani *et al*., 2012 and Ensembl, 2024.

Appendix Table 22: COG classifications of the SGH and AGH protein-coding genes with variants

COG category	Function	SVs	SNPs	INDELs
Information storage and processing				
A	RNA processing and modification	22	40	308
B	Chromatin Structure and dynamics	7	14	133
J	Translation	15	22	202
K	Transcription	97	160	997
L	Replication and repair	273	227	593
Cellular processes and signalling				
D	Cell cycle control and mitosis	15	27	193
M	Cell wall/membrane/envelop biogenesis	2	9	59
N	Cell motility	1	1	6
O	Post-translational modification, protein turnover, chaperone functions	131	190	900
T	Signal Transduction	202	432	2398
U	Intracellular trafficking and secretion	37	47	453
V	Defence mechanism	6	26	119
W	Extracellular structures	15	52	198
X	Mobilome	0	0	0
Y	Nuclear structure	2	3	39
Z	Cytoskeleton	26	66	443
Metabolism				
C	Energy production and conversion	17	36	153
E	Amino Acid metabolism and transport	62	384	922
F	Nucleotide metabolism and transport	11	20	105
G	Carbohydrate metabolism and transport	22	74	363
H	Coenzyme metabolism	1	6	56
I	Lipid metabolism	21	48	267
P	Inorganic ion transport and metabolism	22	56	340
Q	Secondary Structure	8	19	131
Poorly characterised				
R	General function prediction only	0	0	0
S	Function Unknown	843	1,974	5,499
Undetermined		17	30	71
	Total	**1,875**	**3,963**	**14,948**

Appendix Table 23: Protein-coding genes with ³ 15 high-impact variants with functions in KEGG pathways

Variant type	Predicted protein	Locus tag	COG category	KEGG KO term
	Anaphase-promoting complex subunit 4 WD40 domain	CM020037.1.g4.t1	DO	K03364
	Class II histocompatibility antigen, B-L beta	CM020049.1.g2.t1	S	K06752
	Fizzy-related protein	CM020055.1.g628.t1	DO	K03364
	Cadherin repeats.	CM020064.1.g1243.t1	S	K16507
SVs	Class II histocompatibility antigen, B-L beta	WNMG01000082.1.g17.t1	S	K06752
	apolipoprotein	CM020024.1.g1241.t1	I	K14462
	Reverse transcriptase (RNA-dependent DNA polymerase)	CM020025.1.g469.t1	O	K10577
	huntingtin	CM020026.1.g344.t1	S	K04533
	Creveld syndrome	CM020026.1.g417.t1	S	K19605
	UDP-glucose 6-dehydrogenase	CM020026.1.g757.t1	GT	K00012
	Ras responsive element binding protein 1	CM020027.1.g181.t1	K	K20210
	ribonuclease P	CM020027.1.g241.t1	S	K14530
	C2H2-type zinc finger	CM020028.1.g235.t1	S	K09228
	Domain first found in C1r, C1s, uEGF, and bone morphogenetic protein.	CM020028.1.g548.t1	T	K14616
	Glycosyl hydrolase family 1	CM020029.1.g120.t1	G	K01229
	signal transducer and activator of transcription 4	CM020029.1.g634.t1	T	K11222
	Collagen alpha-1(VI) chain	CM020029.1.g706.t1	W	K06238
	Collagen alpha-1(XVII) chain	CM020032.1.g280.t1	W	K07603
	Steroid 17-alpha-hydroxylase 17,20	CM020032.1.g301.t1	Q	K00512
	Chondroitin sulfate N-acetylgalactosaminyltransferase 2	CM020032.1.g729.t1	G	K00746
	Collagen alpha-1(XIV) chain	CM020033.1.g706.t1	W	K08133
	SMG7 nonsense mediated mRNA decay factor	CM020034.1.g7.t1	A	K14409
	26S proteasome non-ATPase regulatory subunit 2	CM020035.1.g224.t1	O	K03028
	segment polarity protein	CM020035.1.g234.t1	S	K02353
	Fanconi anemia group M protein	CM020036.1.g27.t1	L	K10896
	Actin-related protein 8	CM020037.1.g86.t1	Z	K11673
	Fanconi anemia group I protein	CM020038.1.g359.t1	S	K10895
	Vacuolar protein	CM020038.1.g526.t1	U	K20182
	Phosphorylase b kinase regulatory subunit	CM020039.1.g4.t1	G	K07190
	regulator of nonsense transcripts	CM020039.1.g159.t1	A	K14328
Small INDELs	subunit 12	CM020039.1.g343.t1	K	K15162
	CREB-binding protein	CM020044.1.g109.t1	K	K04498
	Unconventional myosin-XV	CM020044.1.g391.t1	Z	K10361
	Transcription initiation factor TFIID subunit 4	CM020047.1.g158.t1	K	K03129
	adenosylhomocysteinase	CM020047.1.g461.t1	H	K01251
	uracil-DNA glycosylase	CM020048.1.g238.t1	L	K03648
	DEP domain-containing protein 5	CM020048.1.g293.t1	T	K20404
	Belongs to the intermediate filament family	CM020049.1.g361.t1	S	K07604

	histone deacetylase	CM020049.1.g489.t1	B	K11406
	PAB-dependent poly(A)-specific ribonuclease subunit PAN2	CM020050.1.g396.t1	L	K12571
	NCK-associated protein 1-like	CM020050.1.g486.t1	S	K05750
	fatty acid synthase	CM020051.1.g8.t1	I	K00665
	intermediate chain 2	CM020051.1.g89.t1	Z	K11143
	exocyst complex component 7	CM020051.1.g267.t1	U	K07195
	Unc-13 homolog A	CM020055.1.g634.t1	TU	K15293
	Histone-lysine	CM020058.1.g225.t1	K	K09186
	DC-STAMP domain-containing protein 2	CM020061.1.g150.t1	S	K14315
	retinol dehydrogenase	CM020062.1.g49.t1	Q	K11150
	Solute carrier family 8	WNMG01000093.1.g16.t1	PT	K05849

Appendix Table 24: Protein-coding genes with ³ 15 high-impact variants undergoing positive and negative selection

Orthogroup	Abyssinian Ground Hornbill Locus tag	Southern Ground Hornbill Locus tag	dN/dS sliding graph	COG category	Predicted protein	KEGG KO term
dN/dS < 1.00						
OG0024315	WNMG01000112.1.g1.t1	Scaff00102_Aln_WNMG01000112.1.g2.t1	0	N/A	N/A	N/A
OG0013540	CM020028.1.g235.t1	Scaff00005_Aln_CM020028.1.g234.t1	0,1473	S	C2H2-type zinc finger	K09228
OG0019861	CM020051.1.g117.t1	Scaff00028_Aln_CM020051.1.g109.t1	0,1694	I	cassette subfamily A member	K05649, K05650, K05651, K05652
OG0003358	CM020026.1.g910.t1	Scaff00003_Aln_CM020026.1.g884.t1	0,1834	S	kiaa1211	-
OG0010333	CM020029.1.g120.t1	Scaff00006_Aln_CM020029.1.g126.t1	0,1901	G	Glycosyl hydrolase family 1	K01229
OG0007411	CM020046.1.g87.t1	Scaff00023_Aln_CM020046.1.g81.t1	0,194	M	ZU5 domain	K10380, K21440
OG0003035	CM020039.1.g393.t1	Scaff00016_Aln_CM020039.1.g383.t1	0,2087	T	Dedicator of cytokinesis	K21853
OG0012775	CM020050.1.g436.t1	Scaff00027_Aln_CM020050.1.g436.t1	0,2608	K	PHD-finger	K09187, K09188
OG0011025	CM020032.1.g157.t1	Scaff00009_Aln_CM020032.1.g165.t1	0,2681	S	PDZ domain-containing protein 8	-
OG0003675	CM020024.1.g1241.t1	Scaff00001_Aln_CM020024.1.g1201.t1	0,2824	I	apolipoprotein	K14462
OG0005057	CM020035.1.g224.t1	Scaff00012_Aln_CM020035.1.g224.t1	0,2882	O	26S proteasome non-ATPase regulatory subunit 2	K03028
OG0010626	CM020030.1.g487.t1	Scaff00007_Aln_CM020030.1.g493.t1	0,3245	S	SON DNA binding protein	-
OG0021335	CM020055.1.g634.t1	Scaff00032_Aln_CM020055.1.g630.t1	0,3493	TU	Unc-13 homolog A	K15293
OG0010008	CM020063.1.g1706.t1	Scaff00040_Aln_CM020063.1.g1329.t1	0,4249	T	Sushi, von Willebrand factor type A, EGF and pentraxin	K17495
OG0004754	CM020024.1.g3062.t1	Scaff00001_Aln_CM020024.1.g2991.t1	0,4779	T	Protocadherin Fat 3	K16506
OG0015071	CM020038.1.g526.t1	Scaff00015_Aln_CM020038.1.g511.t1	0,4967	U	Vacuolar protein	K20182
OG0008685	CM020028.1.g271.t1	Scaff00005_Aln_CM020028.1.g272.t1	0,5181	U	Domain of unknown function (DUF4800)	-
OG0001605	CM020063.1.g1363.t1	Scaff00040_Aln_CM020063.1.g1091.t1	0,5206	PT	receptor 98	K18263
OG0003518	CM020027.1.g1027.t1	Scaff00004_Aln_CM020027.1.g981.t1	0,5582	U	WD domain, G-beta repeat	K14299, K16725
OG0017135	CM020044.1.g109.t1	Scaff00021_Aln_CM020044.1.g108.t1	0,5685	K	CREB-binding protein	K04498
OG0009617	CM020029.1.g492.t1	Scaff00006_Aln_CM020029.1.g470.t1	0,5732	Z	Immunoglobulin	K12567
OG0021331	CM020055.1.g628.t1	Scaff00014_Aln_CM020037.1.g1.t1	0,5981	DO	Fizzy-related protein	K03364

OG0005638	CM020025.1.g469.t1	Scaff00002_Aln_CM020025.1.g449.t1	0,6035	O	Reverse transcriptase (RNA-dependent DNA polymerase)	K10577
OG0005835	CM020039.1.g159.t1	Scaff00016_Aln_CM020039.1.g158.t1	0,6305	A	regulator of nonsense transcripts	K14328
OG0008741	CM020026.1.g1765.t1	Scaff00003_Aln_CM020026.1.g1718.t1	0,6398	S	Galactose binding lectin domain	K04594
OG0013615	CM020029.1.g78.t1	Scaff00006_Aln_CM020029.1.g81.t1	0,6456	S	glycosyltransferase-like domain-containing protein 1	-
OG0009668	CM020027.1.g181.t1	Scaff00004_Aln_CM020027.1.g180.t1	0,6478	K	Ras responsive element binding protein 1	K20210
OG0003250	CM020047.1.g324.t1	Scaff00024_Aln_CM020047.1.g336.t1	0,6501	A	Helicase associated domain (HA2) Add an annotation	K13117
OG0003745	CM020032.1.g292.t1	Scaff00009_Aln_CM020032.1.g306.t1	0,6549	A	Programmed cell death 11	K14792
OG0018008	CM020047.1.g14.t1	Scaff00024_Aln_CM020047.1.g14.t1	0,6557	S	Cadherins are calcium-dependent cell adhesion proteins	K06812
OG0002875	CM020033.1.g706.t1	Scaff00010_Aln_CM020033.1.g679.t1	0,6618	W	Collagen alpha-1(XIV) chain	K08133
OG0005772	CM020039.1.g6.t1	Scaff00016_Aln_CM020039.1.g5.t1	0,6825	S	Apx/Shroom domain ASD2	K18625
OG0009133	CM020055.1.g568.t1	Scaff00032_Aln_CM020055.1.g560.t1	0,69	P	Belongs to the cation transport ATPase (P-type) (TC 3.A.3) family. Type IV subfamily	K01530
OG0007518	CM020047.1.g102.t1	Scaff00024_Aln_CM020047.1.g100.t1	0,6974	A	Zinc finger CCCH-type with G patch	-
OG0019626	CM020050.1.g396.t1	Scaff00027_Aln_CM020050.1.g397.t1	0,701	L	PAB-dependent poly(A)-specific ribonuclease subunit PAN2	K12571
OG0008504	CM020026.1.g344.t1	Scaff00003_Aln_CM020026.1.g335.t1	0,7014	S	huntingtin	K04533
OG0009512	CM020059.1.g24.t1	Scaff00036_Aln_CM020059.1.g24.t1	0,7051	K	Histone-lysine N-methyltransferase NSD3	K07117, K11424, K11425, K15588
OG0019715	CM020050.1.g525.t1	Scaff00027_Aln_CM020050.1.g509.t1	0,7123	W	Collagen alpha-1(II) chain	K19719, K19720
OG0007636	CM020047.1.g461.t1	Scaff00024_Aln_CM020047.1.g470.t1	0,7164	H	adenosylhomocysteinase	K01251
OG0001976	CM020063.1.g805.t1	Scaff00040_Aln_CM020063.1.g656.t1	0,7261	O	cardiomyopathy-associated protein 5	-
OG0014251	CM020044.1.g4.t1	Scaff00021_Aln_CM020044.1.g7.t1	0,7351	S	Rogdi homolog	-

OG0006678	CM020025.1.g1421.t1	Scaff00002_Aln_CM020025.1.g1332.t1	0,7428	A	Belongs to the DEAD box helicase family	K14777
OG0003005	CM020038.1.g114.t1	Scaff00015_Aln_CM020038.1.g112.t1	0,7536	S	chondroitin sulfate proteoglycan 4	K08115
OG0014193	CM020036.1.g646.t1	Scaff00013_Aln_CM020036.1.g635.t1	0,7546	V	YLP motif-containing protein	K17602
OG0010436	CM020029.1.g493.t1	Scaff00006_Aln_CM020029.1.g471.t1	0,7615	T	SEC14 domain and spectrin repeat-containing protein 1	-
OG0004752	CM020034.1.g215.t1	Scaff00011_Aln_CM020034.1.g206.t1	0,7625	S	Translocated promoter region, nuclear basket protein	K09291, K20478
OG0004681	CM020034.1.g7.t1	Scaff00011_Aln_CM020034.1.g6.t1	0,763	A	SMG7 nonsense mediated mRNA decay factor	K14409
OG0004538	CM020033.1.g178.t1	Scaff00010_Aln_CM020033.1.g173.t1	0,7668	S	zinc finger and BTB	K10497
OG0007226	CM020044.1.g391.t1	Scaff00021_Aln_CM020044.1.g394.t1	0,7669	Z	Unconventional myosin-XV	K10361
OG0011060	CM020032.1.g280.t1	Scaff00009_Aln_CM020032.1.g291.t1	0,7708	W	Collagen alpha-1(XVII) chain	K07603
OG0003609	CM020028.1.g321.t1	Scaff00005_Aln_CM020028.1.g315.t1	0,7796	T	Obscurin, cytoskeletal calmodulin and titin-interacting RhoGEF	K17531
OG0003669	CM020029.1.g634.t1	Scaff00006_Aln_CM020029.1.g604.t1	0,7796	T	signal transducer and activator of transcription 4	K11222
OG0017998	CM020047.1.g3.t1	Scaff00024_Aln_CM020047.1.g1.t1	0,7798	N/A	N/A	N/A
OG0011069	CM020032.1.g301.t1	Scaff00009_Aln_CM020032.1.g315.t1	0,7826	Q	Steroid 17-alpha-hydroxylase 17,20	K00512
OG0013278	CM020035.1.g138.t1	Scaff00012_Aln_CM020035.1.g141.t1	0,7832	T	TRAF2 and	K04407, K08840
OG0008320	CM020051.1.g255.t1	Scaff00028_Aln_CM020051.1.g251.t1	0,7848	O	RING finger protein unkempt homolog	-
OG0008223	CM020051.1.g8.t1	Scaff00028_Aln_CM020051.1.g8.t1	0,7856	I	fatty acid synthase	K00665
OG0015341	CM020039.1.g343.t1	Scaff00016_Aln_CM020039.1.g336.t1	0,7879	K	subunit 12	K15162
OG0003030	CM020039.1.g303.t1	Scaff00016_Aln_CM020039.1.g298.t1	0,7883	S	FERM and PDZ	-
OG0003012	CM020038.1.g359.t1	Scaff00015_Aln_CM020038.1.g339.t1	0,7907	S	Fanconi anemia group I protein	K10895
OG0008327	CM020051.1.g267.t1	Scaff00028_Aln_CM020051.1.g264.t1	0,7996	U	exocyst complex component 7	K07195
OG0010890	CM020031.1.g604.t1	Scaff00008_Aln_CM020031.1.g567.t1	0,8121	U	exocyst complex component 8	K19986
OG0024242	WNMG01000093.1.g16.t1	Scaff00084_Aln_WNMG01000093.1.g11.t1	0,8121	PT	Solute carrier family 8	K05849
OG0007500	CM020047.1.g51.t1	Scaff00024_Aln_CM020047.1.g50.t1	0,8163	S	DNA (cytosine-5-)-methyltransferase	K17398, K17399
OG0004815	CM020034.1.g392.t1	Scaff00011_Aln_CM020034.1.g394.t1	0,8261	S	Zinc finger protein 644	-

OG0007751	CM020048.1.g238.t1	Scaff00025_Aln_CM020048.1.g236.t1	0,8321	L	uracil-DNA glycosylase	K03648
OG0009675	CM020027.1.g241.t1	Scaff00004_Aln_CM020027.1.g242.t1	0,8427	S	ribonuclease P	K14530
OG0008522	CM020026.1.g417.t1	Scaff00003_Aln_CM020026.1.g407.t1	0,8437	S	Creveld syndrome	K19605
OG0012608	CM020061.1.g149.t1	Scaff00038_Aln_CM020061.1.g129.t1	0,8496	S	DC-STAMP domain-containing protein	K22375
OG0015541	CM020040.1.g42.t1	Scaff00017_Aln_CM020040.1.g49.t1	0,8538	S	axonemal central apparatus assembly	K17570
OG0008576	CM020026.1.g757.t1	Scaff00003_Aln_CM020026.1.g727.t1	0,8563	GT	UDP-glucose 6-dehydrogenase	K00012
OG0007547	CM020047.1.g158.t1	Scaff00024_Aln_CM020047.1.g165.t1	0,8645	K	Transcription initiation factor TFIID subunit 4	K03129
OG0006272	CM020041.1.g474.t1	Scaff00018_Aln_CM020041.1.g481.t1	0,8737	S	Cadherin C-terminal cytoplasmic tail, catenin-binding region	K16496, K16497, K16499
OG0017368	CM020044.1.g383.t1	Scaff00021_Aln_CM020044.1.g386.t1	0,8832	S	Retinoic acid-induced protein 1	K19749
OG0024092	WNMG01000077.1.g8.t1	Scaff00068_Aln_WNMG01000077.1.g9.t1	0,8954	N/A	N/A	N/A
OG0002274	CM020025.1.g227.t1	Scaff00002_Aln_CM020025.1.g209.t1	0,8971	D	Protein MCM10 homolog	K10736
OG0001146	CM020025.1.g1316.t1	Scaff00002_Aln_CM020025.1.g1233.t1	0,9033	S	protein localization to centrosome	K16547
OG0024129	WNMG01000078.1.g38.t1	Scaff00069_Aln_WNMG01000078.1.g34.t1	0,9184	E	FACT complex subunit SPT16	-
OG0013277	CM020035.1.g137.t1	Scaff00012_Aln_CM020035.1.g139.t1	0,9281	G	Solute carrier family 2, facilitated glucose transporter member 2	K07299, K07593
OG0019693	CM020050.1.g490.t1	Scaff00027_Aln_CM020050.1.g484.t1	0,9314	W	Integrin alpha	K06484, K06487, K06584
OG0013863	CM020032.1.g845.t1	Scaff00009_Aln_CM020032.1.g854.t1	0,938	S	Secreted protein acidic and rich in cysteine Ca binding region	K08136
OG0016727	CM020057.1.g248.t1	Scaff00034_Aln_CM020057.1.g253.t1	0,9478	S	Cell-cycle sustaining, positive selection,	-
OG0024303	WNMG01000103.1.g6.t1	Scaff00094_Aln_WNMG01000103.1.g6.t1	0,952	O	ATPase family AAA domain-containing protein	-
OG0005224	CM020036.1.g27.t1	Scaff00013_Aln_CM020036.1.g24.t1	0,9657	L	Fanconi anemia group M protein	K10896
OG0007967	CM020049.1.g361.t1	Scaff00026_Aln_CM020049.1.g352.t1	0,9734	S	Belongs to the intermediate filament family	K07604
OG0003312	CM020051.1.g89.t1	Scaff00028_Aln_CM020051.1.g80.t1	0,9836	Z	intermediate chain 2	K11143

dN/dS > 1.00						
OG0008944	CM020055.1.g11.t1	Scaff00032_Aln_CM020055.1.g15.t1	1,0135	T	Receptor-type tyrosine-protein phosphatase V-like	K18036
OG0024040	CM020063.1.g2830.t1	Scaff00040_Aln_CM020063.1.g2208.t1	1,041	C	Retinal dehydrogenase 1	K00128, K07249
OG0009227	CM020056.1.g191.t1	Scaff00033_Aln_CM020056.1.g188.t1	1,0545	I	phytanoyl-CoA dioxygenase domain-containing protein 1 isoform X1	K00477, K18741
OG0002220	CM020037.1.g86.t1	Scaff00014_Aln_CM020037.1.g91.t1	1,0644	Z	actin-related protein 8	K11673
OG0013933	CM020061.1.g150.t1	Scaff00038_Aln_CM020061.1.g130.t1	1,1731	S	DC-STAMP domain-containing protein 2	K14315
OG0008567	CM020026.1.g731.t1	Scaff00003_Aln_CM020026.1.g699.t1	1,1809	S	NACHT and WD repeat	-
OG0016172	CM020041.1.g353.t1	Scaff00018_Aln_CM020041.1.g365.t1	1,3555	O	WW and C2 domain containing 1	K16685