scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Whole genome sequencing, assembly and annotation of the Southern Ground Hornbill – *Bucorvus leadbeateri*

Jasmin Patel, Angela Botes, Jean Mollett & Pieter De Maayer ✉

The Southern Ground Hornbill (SGH – *Bucorvus leadbeateri*) is one of the largest hornbill species worldwide, known for its complex social structures and breeding behaviours. This bird has been of great interest due to its declining population and disappearance from historic ranges in southern Africa. Despite being the focus of numerous conservation efforts, with research forming an integral part of these initiatives, there is still a substantial lack of knowledge regarding the molecular biology aspects of this bird species. In this study, whole genome sequencing of the SGH was achieved using Illumina short-read (NovaSeq 6000) and Pacific Biosciences long-read technologies. A hybrid *de novo* genome assembly followed by reference-based refinement produced a 1.16 Gb high-quality draft genome assembly of the SGH comprised of 1,672 contigs (N50 value of 40.45 Mb). 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.

Background & Summary

The Southern Ground Hornbill (SGH – *Bucorvus leadbeateri*), belonging to the family *Bucorvidae* and the avian order Bucerotiformes, is the largest hornbill species in the world[1,2]. It is one of the only two species classified under the genus *Bucorvus* and is distributed across sixteen African countries[3–5]. These birds are distinct from other hornbill species in the order *Bucerotiformes* because they do not completely seal their females in nest cavities during breeding[6]. Additionally, the SGH is globally known as the largest avian cooperative breeder, primarily maintaining monogamous relationships[1]. These birds have complex social structures, with individuals occurring in groups of two to eleven individuals that include the main breeding pair (alpha male and female), juveniles and non-breeding adult members, with the latter typically being helper males[7]. The SGH hold significant ecological and cultural importance in their environment[4,8]. Moreover, with the groups being largely territorial and occupying year-round territories, these birds are considered a flagship conservation species of the savannah and grassland biomes[5].

The SGH has been globally classified as Vulnerable by the International Union for Conservation of Nature (IUCN) and regionally as Endangered within the southern extent of its dispersal range[9–12]. In South Africa, the SGH has been estimated to have lost 70% of its range and 50% of its total historical population[13,14], and between 1,290 and 2,380 individuals are estimated to reside in protected areas[11]. The decline in populations has been largely attributed to the afforestation and bush encroachment of grasslands and savannahs, respectively, resulting in the loss of suitable habitats and nesting sites[5]. Furthermore, local threats to populations have been linked to persecution for window breaking, indirect poisoning, the aviculture trade and traditional cultural beliefs[5,14,15].

Conservation efforts to safeguard the remaining populations of the SGH have played a significant role in improving our understanding of this bird's lifestyle and behavioural patterns. However, limited data on its molecular biology are available, aside from a restricted set of microsatellite markers and the complete mitochondrial genome. These resources have mainly been used in phylogenetic and population dynamic analysis[16,17]. In this context, whole genome sequencing of the SGH will provide insight into these birds' genetic composition and potential. This information will aid in identifying key genes involved in the evolution and adaptation of

School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, 2017, South Africa. ✉e-mail: pieter.demaayer@wits.ac.za

Assembly	Illumina (PE)	CLR reads	HiFi reads
Number of reads (million)	45.15	5.09	0.33
Average read length (bp)	290 (2 × 145)	6,949	7,847
Total length of reads (Gb)	13.07	35.36	2.57
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

Table 1. Genome sequencing statistics of the PacBio and Illumina reads.

these birds to their environment through comparative genomic analysis with other hornbill species and offer an avenue to improve current conservation efforts.

The genome of the SGH was sequenced using paired-end 2 × 150 bp Illumina short-read sequencing, which produced 13.07 Gb or 45.15 million paired-end reads, and PacBio long-read sequencing which produced 35.36 Gb or 5.09 million CLR and 2.57 Gb or 0.33 million HiFi reads (Table 1). A hybrid *de novo* assembly approach using DBG2OLC[18] was employed (Fig. 1), resulting in a 1.12 Gb assembly comprised of 10,518 contigs (largest contig: 2.06 Mb), with an N50 value of 0.26 Mb and a genome coverage fraction of 95.48% when mapped against the Abyssinian Ground Hornbill genome (AGH – *Bucorvus abyssinicus*; GCA_009769605.1) (Table 2). These two *Bucorvus* species are estimated to have diverged from a common ancestor approximately 2.52 million years ago, during the late Pliocene to early Pleistocene[4]

Following contig extension, reference-guided scaffolding using the AGH genome, gap closing and polishing, the final high-quality SGH draft genome was 1.16 Gb in size and comprised 1,672 scaffolds with an N50 value of 40.45 Mb (Table 2**;** Table S1). A total of 1.10 Gb (90.38%) of the SGH assembly aligned with the genome assembly of the AGH (Table S1). This included 41 scaffolds that aligned to chromosomes 1–39 and the sex chromosomes W and Z of the AGH. Additionally, 61 out of 71 unplaced scaffolds of the AGH assembly had matches in the SGH assembly. A further 1,410 contigs partially aligned to the AGH scaffolds, while 161 contigs (average length: 15,039 bp, largest contig: 0.22 Mb) did not align with the AGH genome.

The SGH draft genome was structurally and functionally annotated using MOSGA[19] in an iterative process. Augustus v 3.5.0[20] trained with the gene models of the chicken and subsequently refined with the BUSCO v 5.6.1[21] output, predicted 33,746 proteins. A total of 28,825 of the predicted proteins were functionally annotated using eggNOG-mapper v2[22] (Fig. 2).

A total of 88.60 Mb sequences were masked as repetitive elements which constituted ~7.64% of the draft SGH genome assembly, with ~59.09 Mb of the elements comprised of class I retrotransposons which included the long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Table 3). Furthermore, 4.00% (3.62 Mb) of these interspersed repeats were unclassified. A total of 280 tRNAs were predicted on the draft genome using tRNAscan-SE v 2.0[23] of which 162 decode the standard 20 amino acids. Moreover, 50 rRNAs were predicted with Barrnap[24], which were predominated by the eukaryotic 28S (17 subtypes) and 5S (15 subtypes) rRNA subtypes.

The completeness of the final SGH draft genome was assessed using BUSCO v 5.6.1[21] with the Avian lineage dataset (Aves_odb10) comprising of 8,338 BUSCOs. The SGH genome assembly incorporates 7,951 (95.4%) complete BUSCOs (7,910 single-copy and 41 duplicated BUSCOs) and 85 fragmented BUSCOs, while 302 BUSCOs were missing.

Methods

Sample collection and preparation. A blood sample was collected from a female SGH in captivity from the Montecasino Bird Garden, Johannesburg, South Africa in collaboration with the Mabula Ground Hornbill Project. Ethics clearance was obtained from the University of the Witwatersrand Animal Ethics Screening Committee (Clearance number: 2020/09/03B). Total genomic DNA was extracted using a 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). 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 50 µl of TE buffer (0.01 M Tris-HCl and 0.001 M EDTA). DNA concentration and quality was evaluated using a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, Massachusetts, USA). The gDNA was shipped to Molecular Research LP (www.mrdnalab.com, Shallowater, Texas, USA) for long- and short-read sequencing.

Library preparation and sequencing. *PacBio Sequel library preparation and sequencing.* The gDNA was sheared using the 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 Biosciences,

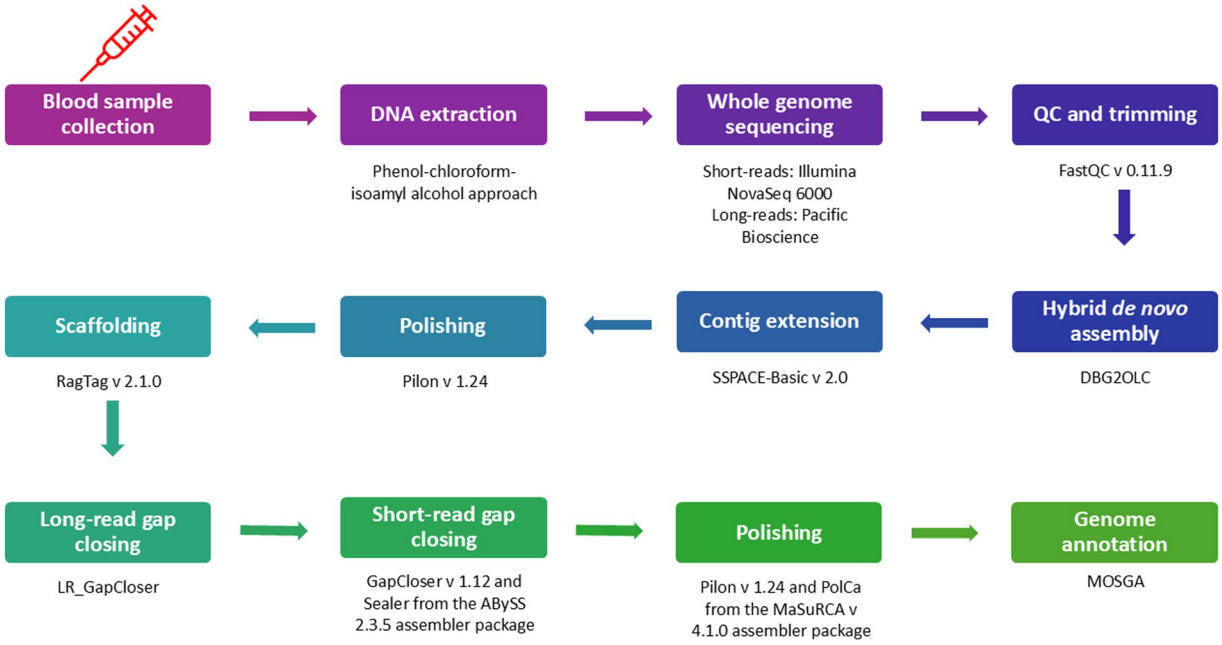

Fig. 1 Schematic representation of the whole genome sequencing, assembly and annotation of the Southern Ground Hornbill.

Assembly	DBG2OLC assembly	Final refined assembly
Number of contigs/scaffolds	10,518 contigs	1,672 scaffolds
Largest contig (Mb)	2.06	137.99
Total length of genome (Gb)	1.12	1.16
Reference genome length (Gb)	1.13	1.13
GC (%)	43.50	43.50
N50 (Mb)	0.26	40.45
N90 (Mb)	0.05	13.42
L50 (bp)	1,289	9
L90 (bp)	4,959	31
Genome fraction (%)	95.47	96.29

Table 2. Assembly statistics of the refined SGH draft genome assembly.

California, USA). During library preparation, single-strand overhangs were removed from the sheared (fragmented) DNA, and the fragmented ends and DNA damage were repaired using DNA repair enzymes available in the SMRTbell Express Template Prep kit 2.0 (Pacific Biosciences, California, USA). Following library preparation, DNA fragments that ranged between 6 to 10 kilobase (kb) pairs were selected using the Blue Pippin automated size-selection instrument (Sage Science, Massachusetts, USA). The final concentration of the library was evaluated using 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 platform (Pacific Biosciences, California, USA), using the paired-end 10-hour sequencing protocol. This produced Continuous Long Reads (CLR) and High Fidelity (HiFi) reads, which were combined and used for genome assembly and subsequent analysis.

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). The genomic DNA sample was diluted, tagged with adapters and simultaneously fragmented. Following library preparation, the final concentration of the library was evaluated using 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 NovaSeq6000 platform (Illumina, California, USA), following a paired-end 300 cycle protocol.

Hybrid *de novo* assembly. The quality of the raw PacBio and Illumina reads was evaluated using FastQC v 0.11.9[25]. 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[26]. Hybrid *de novo* assembly was performed using

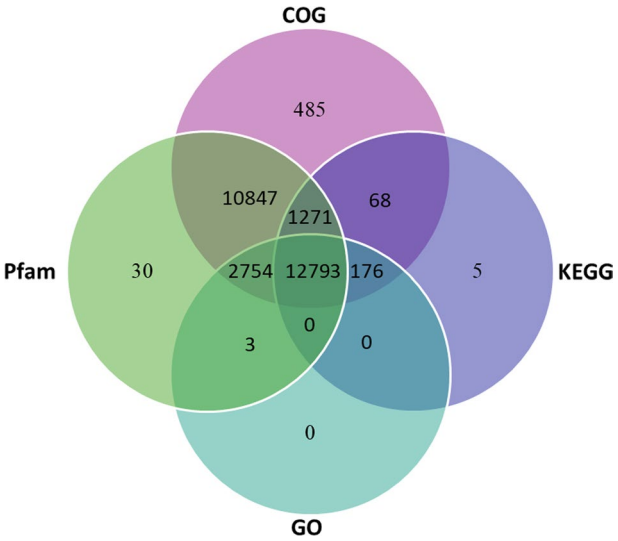

Fig. 2 Functional classification of annotated genes in the Southern Ground Hornbill genome by different databases.

Types of repeats	Class	Number of elements	% of sequence
Retroelements		**186,729**	65.40
	SINEs:	52,684	7.28
	Penelope	276	0.06
	LINEs:	131,842	57.84
	L2/CR1/Rex	131,309	57.71
	L1/CIN4	257	0.06
	LTR	2,203	0.28
	Gypsy/DIRS1	313	0.07
DNA transposons		**103,581**	12.90
	hobo-Activator	25,382	2.75
	Tc1-IS630-Pogo	4,553	0.51
	Tourist/Harbinger	26,386	4.09
Rolling-circles		**588**	0.07
Unclassified:		**47,596**	4.00
Total interspersed repeats			**82.30**
Small RNA		**70,501**	5.29
Satellites		**0**	0.00
Simple repeats		**205,915**	10.20
Low complexity		**40,633**	2.20
Total		**655,543**	**100.00**

Table 3. Relative proportion of repetitive elements within the SGH genome.

DBG2OLC[18]. In brief, the trimmed Illumina reads were first assembled using SparseAssembler using the optimal parameters: k-mer size = 31, gap size = 13, NodeCovTh = 1 and EdgeCovTh = 0. The Illumina assembly was then mapped against the PacBio sequence and a consensus sequence was generated using basic local alignment with successive refinement (BLASR) v 5.3.5[27] and Sparc[28]. The following optimal parameters were used during these stages: AdaptiveTh = 0.003, KmerCovTh = 2 and MinOverlap = 21.

Contig extension was undertaken on the resultant assembly using SSPACE-Basic v 2.0[29] (with default parameters) with the raw Illumina and PacBio HiFi reads. To improve the quality of the draft assembly, polishing was conducted using ten iterations of Pilon v 1.24[30]. The polished contigs were then scaffolded against the AGH genome assembly (GCA_009769605.1) using RagTag v 2.1.0[31]. Gap closing using the PacBio reads was performed using LR_GapCloser[32]. Two iterations of gap closing using the Illumina reads were performed using first GapCloser v 1.12[33] and then Sealer from the ABySS v 2.3.5 assembler package[34,35] on the resultant output. A final round of polishing was conducted using five iterations of Pilon and then five iterations of Polishing by Calling Alternatives (POLCA) from the MaSuRCA v 4.1.0. assembler package[36]. Quast-LG v 5.2.0[37] was systematically implemented to evaluate the quality of the refined assembly at each stage.

Genome annotation. Genome annotation was performed using the annotation tool Modular Open-Source Genome Annotator (MOSGA)[19]. Repetitive elements were identified in the genome using RepeatMasker[38] with the chicken (*Gallus gallus*) repeat model and subsequently soft-masked. *Ab initio* gene prediction was performed using Augustus[20] trained with the gene models of the chicken (GalGal4). Gene prediction using Augustus v 3.5.0 was further refined using the Benchmarking Universal Single-Copy Orthologs (BUSCO) v 5.6.1[21] output. Predicted proteins were annotated using evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG)-mapper v2[22] against the eggNOG 5.0 database[39] with default parameters. eggNOG functionally annotated the predicted genes using the Conserved Orthologous Groups (COGs) categories[40], Gene ontology (GO) terms[41,42], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways[43], Simple Modular Architecture Research Tool (SMART)/PFAM domains[44,45], Carbohydrate-active enzymes (CAZy)[46] and KEGG modules[47].

Noncoding RNA elements were predicted using tRNAscan-SE v 2.0[23] (with default parameters) and Barrnap v 0.9[24] (set to identify eukaryotic 5S, 5.8S, 28S and 18S with a minimum E-value of 10e-6) to predict transfer RNAs and ribosomal RNAs, respectively. Completeness of the final draft genome and the quality of the predicted gene models was assessed using BUSCO v. 5.6.1[21] with default parameters against the Aves lineage dataset aves_odb10 (8,338 BUSCOs).

Data Records

The raw Illumina and PacBio Hifi and raw sequencing data for the SGH genome have been deposited in the NCBI Sequence Read Archive (SRA) database under accession number SRP547136[48]. The final assembly is available in the GenBank database under accession number JBEWAY000000000[49].

Technical Validation

To ensure the integrity of the genomic research, the initial step involved evaluating the quality of the raw sequenced data with FastQC. Following this, the Quast-LG software application was implemented to assess the assembled data. During the refinement steps, Quast-LG was further used to evaluate the contiguity, quality, and completeness of the SGH genome, using the AGH genome as a reference. This evaluation highlighted the impact of each refinement step on the SGH genome assembly metrics, which in turn facilitated further optimization of parameters where required. It also guided the selection of the most effective software applications for each refinement step. The completeness of the genome assembly and the quality of the annotations were assessed using BUSCO implemented in MOSGA. This approach allowed for the systematic refinement of the SGH genome, ensuring its quality and reliability for subsequent scientific inquiry.

Usage Notes

The authors present no control or limitations to the data.

Ethics declaration. All animal procedures including blood sample collection from the female Southern Ground Hornbill, a threatened and protected species in South Africa, was obtained following approval by the following committees and departments: Ethics clearance from the University of the Witwatersrand Animal Ethics Screening Committee (Clearance number: 2020/09/03B). Clearance from the Department of Agriculture, Land Reform and Rural Development (DALRRD) in South Africa, 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)). Clearance from the Department of Environmental Affairs in South Africa, Threatened Or Protected Species (TOPS) permit to work on a threatened or protected species (Certificate number: 29480).

Code availability

All software applications and bioinformatic pipelines utilised in this study were implemented according to the published user protocols and manuals. Where applicable, software versions and deviation from default setting have been described in the Methods.

Received: 8 October 2024; Accepted: 3 January 2025;
Published online: 11 January 2025

References

1. Kemp, A. Bird families of the world. in *The Hornbills: Bucerotiformes* vol. 114 94–95 (Oxford University Press, Oxford, United Kingdom, 1995).
2. van Vuuren, A. K. J., Kemp, L. V. & McKechnie, A. E. The beak and unfeathered skin as heat radiators in the southern ground-hornbill. *Journal of Avian Biology* **51**, (2020).
3. Kemp, A. C. A review of the hornbills: Biology and radiation. in *The living bird* 105–136 (Cornell Laboratory of Ornithology, Pretoria, South Africa, Transvaal Museum, 1979).
4. Kemp, L. *et al*. Broad-scale genetic assessment of Southern Ground-Hornbills (*Bucorvus leadbeateri)* to inform population management. *Global Ecology and Conservation* **52**, e02963 (2024).
5. Cilliers, D., Evans, S., Coetzee, H. & van Rensburg, L. Developing a site selection tool to assist reintroduction efforts for the Southern Ground-Hornbill Bucorvus leadbeateri. *Ostrich* **84**, 101–111 (2013).
6. Kemp, A. C. & Kemp, M. I. The biology of the Southern Ground Hornbill Bucorvus leadbeateri (Vigors) (Aves: Bucerotidae). *Annals of the Transvaal Museum* **32**, 65–100 (1980).
7. Kemp, A. C. The behavioural ecology of the Southern Ground Hornbill: are competitive offspring at a premium? in *Proceedings of International 100 Deutscheornitologen-Gesselschaft Meeting Current topics in avian biology* vol. 1988 267–271 (Bonn, 1990).
8. Le Roux, J. *The Biodiversity of South Africa 2002: Indicators, Trends and Human Impacts*. (Struik, Cape Town, 2002).
9. BirdLife International. IUCN Red List for birds. (2021).
10. Simmons, R., Brown, C. J. & Kemper, J. (eds). Southern Ground-Hornbill Bucorvus leadbeateri. in *Birds to Watch in Namibia: Red, Rare and Endemic Species* 38–39 (Ministry of Environment and Tourism, Namibia, 2015).

11. Taylor, M. R., Peacock, F. & Wanless, R. M. (eds). Southern Ground Hornbill. in *The 2015 Eskom Red Data Book of Birds of South Africa, Lesotho and Swaziland* 119–121 (BirdLife South Africa, Johannesburg, South Africa, 2015).
12. Kemp, L. V. *et al*. Review of trial reintroductions of the long-lived, cooperative breeding Southern Ground-hornbill. *Bird Conservation International* **30**, 533–558 (2020).
13. Kemp, A. Southern Ground Hornbill. in *In* Hockey, P. A. R., Dean, W. R. J., *and* Ryan, P. G. *(eds). Roberts Birds of Southern Africa* 158–159 (Trustees of the John Voelcker Bird Book Fund, Cape Town, South Africa, 2005).
14. Kemp, A. & Webster, R. *Latest Analysis of Southern Ground Hornbill (SGH) Distribution and Population in South Africa: December 2008*. 1–11 https://docs.google.com/file/d/0B5qf0llRHM7PT2I1bUNDaTRQcXM/edit (2008).
15. Morrison, K. *et al. SouthernGround Hornbill (Bucorvus Leadbeateri) PHVA Workshop Report* https://doi.org/10.13140/RG.2.1.4752.9442 (Conservation Breeding Specialist Group (SSC / IUCN) / CBSG South Africa; Endangered Wildlife Trust, Johannesburg, South Africa, 2005).
16. Delport, W., Ferguson, J. W. H. & Bloomer, P. Characterization and evolution of the mitochondrial DNA control region in hornbills (Bucerotiformes). *J Mol Evol* **54**, 794–806 (2002).
17. Dalton, D. & Kotze, A. Isolation and characterization of microsatellite loci in Southern Ground Hornbills (Bucorvus leadbeateri). (2011).
18. Ye, C., Hill, C. M., Wu, S., Ruan, J. & Ma, Z. S. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Sci Rep* **6**, 31900 (2016).
19. Martin, R., Hackl, T., Hattab, G., Fischer, M. G. & Heider, D. MOSGA: Modular Open-Source Genome Annotator. *Bioinformatics* **36**, 5514–5515 (2020).
20. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465–W467 (2005).
21. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
22. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825–5829 (2021).
23. Chan, P. P. & Lowe, T. M. tRNAscan-SE: Searching for tRNA genes in genomic sequences. *Methods Mol Biol* **1962**, 1–14 (2019).
24. Seemann, T. Barrnap (2013).
25. Andrew, S. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. (2015).
26. H Lab. FASTX-Toolkit. (n.d.).
27. Chaisson, M. J. & Tesler, G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinformatics* **13**, 238 (2012).
28. Ye, C. & Ma, Z. S. Sparc: a sparsity-based consensus algorithm for long erroneous sequencing reads. *PeerJ* **4**, e2016 (2016).
29. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
30. Walker, B. J. *et al*. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* **9**, e112963 (2014).
31. Alonge, M. *et al*. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. *Genome Biology* **23**, 258 (2022).
32. Xu, G.-C. *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**, giy157 (2019).
33. Luo, R. *et al*. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* **1** (2012).
34. Paulino, D. *et al*. Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinformatics* **16**, 230 (2015).
35. Jackman, S. D. *et al*. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. *Genome Res.* **27**, 768–777 (2017).
36. Zimin, A. V. & Salzberg, S. L. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology* **16**, e1007981 (2020).
37. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
38. Smit, A., Hubely, R. & Green, P. RepeatMasker. https://www.repeatmasker.org/ (2021).
39. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* **47**, D309–D314 (2019).
40. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33–36 (2000).
41. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
42. Gene Ontology Consortium *et al*. The Gene Ontology knowledgebase in 2023. *Genetics* **224**, iyad031 (2023).
43. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
44. Schultz, J., Milpetz, F., Bork, P. & Ponting, C. P. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences* **95**, 5857–5864 (1998).
45. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405–420 (1997).
46. Cantarel, B. L. *et al*. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37**, D233–D238 (2009).
47. Kanehisa, M. *et al*. KEGG for linking genomes to life and the environment. *Nucleic Acids Research* **36**, D480–D484 (2008).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP547136 (2024).
49. Patel, J., Botes, A., Mollett, J. & De Maayer, P. Southern Ground Hornbill (bBucLead), whole genome shotgun sequencing project. *Genbank*. https://identifiers.org/ncbi/insdc:JBEWAY00000000 (2024).

Acknowledgements

The authors would like to thank the University of the Witwatersrand Post Graduate Merit Award (PMA) and South African National Research Foundation (NRF) for student funding (JP: NRF PhD Scholarship MND200427517202), the Mabula Ground Hornbill Project, South African National Biodiversity Institute (SANBI) and BirdLife South Africa for funding this research.

Author contributions

J.P., A.B., P.D.M. and J.M. conceived the study. J.P. and P.D.M. performed experiments and analysis, J.P., P.D.M. and J.M. wrote the original manuscript. A.B. provided financial support for the publication of the manuscript. All authors contributed to the final version. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04412-2.

Correspondence and requests for materials should be addressed to P.D.M.