



DATA NOTE

REVISED The genome sequence of the European shag, *Gulosus aristotelis* (previously *Phalacrocorax aristotelis*) (Linnaeus, 1761)

[version 2; peer review: 2 approved with reservations]

Hannah M. Ravenswater¹, Fiona Greco¹, Sarah J. Burthe²,
Emma J. A. Cunningham¹, Darwin Tree of Life Barcoding collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory
team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

¹The University of Edinburgh, Edinburgh, Scotland, UK
²UK Centre for Ecology & Hydrology, Wallingford, England, UK

V2 First published: 11 Mar 2024, 9:144
<https://doi.org/10.12688/wellcomeopenres.21119.1>
Latest published: 22 Sep 2025, 9:144
<https://doi.org/10.12688/wellcomeopenres.21119.2>

Abstract
We present a genome assembly from an individual female *Gulosus aristotelis*, previously known as *Phalacrocorax aristotelis*, (the European shag; Chordata; Aves; Pelecaniformes; Phalacrocoracidae). The genome sequence is 1,279.1 megabases in length. Most of the assembly is scaffolded into 36 chromosomal pseudomolecules, including the Z and W sex chromosomes. Gene annotation of this assembly on Ensembl identified 16,474 protein coding genes. The mitochondrial genome has also been assembled and is 18.61 kilobases in length.

Keywords
Gulosus aristotelis, *Phalacrocorax aristotelis*, European shag, genome sequence, chromosomal, Pelecaniformes



This article is included in the [Tree of Life gateway](#).

Open Peer Review

Approval Status ? ?

	1	2
version 2		
(revision)		
22 Sep 2025		
version 1	?	?
11 Mar 2024	view	view
1. Nicolas J. Rawlence, University of Otago, Dunedin, New Zealand		
2. Xianzhao Kan , Anhui Normal University, Wuhu, China		
Any reports and responses or comments on the article can be found at the end of the article.		

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: **Ravenswater HM:** Conceptualization, Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Greco F:** Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Burthe SJ:** Investigation, Resources, Supervision, Writing – Review & Editing; **Cunningham EJA:** Conceptualization, Funding Acquisition, Investigation, Resources, Supervision, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, <https://doi.org/10.35802/206194>] and the Darwin Tree of Life Discretionary Award [218328, <https://doi.org/10.35802/218328>] and NERC awards NE/S007407/1, NE/L002558/1 and NE/V001779/1.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Ravenswater HM *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Ravenswater HM, Greco F, Burthe SJ *et al.* **The genome sequence of the European shag, *Gulosus aristotelis* (previously *Phalacrocorax aristotelis*) (Linnaeus, 1761) [version 2; peer review: 2 approved with reservations]** Wellcome Open Research 2025, 9:144 <https://doi.org/10.12688/wellcomeopenres.21119.2>

First published: 11 Mar 2024, 9:144 <https://doi.org/10.12688/wellcomeopenres.21119.1>

REVISED Amendments from Version 1

We have corrected the use of *Phalacrocorax aristotelis* to *Gulosus aristotelis* throughout. We have also added a taxonomic note to explain the preferred species name.

We have also revised the results of the Merqury.FK output using the latest version of this software, and given the Earth Biogenome Project summary metric for the assembly.

We have added labels to the Hi-C contact map of the chromosomes in [Figure 5](#).

Any further responses from the reviewers can be found at the end of the article

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Sauropsida; Sauria; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Suliformes; Phalacrocoracidae; *Phalacrocorax*; *Phalacrocorax aristotelis* (Linnaeus, 1761) (NCBI:txid126867).

Taxonomic note. We follow recent avian checklists (e.g. IOC, BirdLife/BOU) that place the European shag in *Gulosus aristotelis* (Linnaeus, 1761), reflecting molecular evidence separating the “shags” from *Phalacrocorax sensu stricto*. NCBI currently retains *Phalacrocorax aristotelis* (NCBI:txid126867); we treat this as a synonym and map our usage of *Gulosus* to the same taxon ID.

Background

Gulosus aristotelis (previously known as *Phalacrocorax aristotelis* and commonly known as the European shag), is a large seabird species within the cormorant family, Phalacrocoracidae (Croxall, 1987; Thanou *et al.*, 2017). European shags have dark-green plumage in adulthood ([Figure 1](#)), but that can often appear black, a yellow gape, green eyes, black legs and feet and a crest during the breeding season (Wanless & Harris, 1997).



Figure 1. Photograph of a *Gulosus aristotelis* from the same study population as the specimen used for genome sequencing. Photograph by Fiona Greco.

Juvenile plumage is lighter brown and present up to two years old (Wanless & Harris, 1997). The species displays sexual dimorphism, with males weighing 1900 g on average, compared to 1600 g in females (Daunt *et al.*, 2001). Sex is typically identified using vocalisations in adults (Snow, 1960) and molecular techniques in juveniles (Thanou *et al.*, 2013).

European shags are benthic foot-propelled pursuit divers, with a partially wettable plumage that facilitates diving in shallow waters but requires individuals to return to shore each day to dry and roost (Grémillet *et al.*, 1998), increasing their vulnerability to inclement weather conditions. The species' diet primarily consists of fish with some predation of crustaceans, with both temporal and spatial variability consumption of prey species (Howells *et al.*, 2017; Howells *et al.*, 2018; Velando & Freire, 1999).

Gulosus aristotelis breeds colonially, creating nest sites on cliff ledges or cavities under rocks (Velando & Freire, 2001; Wanless & Harris, 1997). European shags start breeding at 2 to 3 years (Aebischer, 1986), with the lifespan of over 14 years on average (Herborn *et al.*, 2014), with the oldest individual recorded on the Isle of May, an important seabird breeding colony where shags have been intensively monitored since the 1970s, at 23 years-old (Hall, 2004). Breeding occurs seasonally, with shags laying on average a clutch of three eggs, which hatch asynchronously after an approximately 35-day incubation period (Granroth-Wilding *et al.*, 2014; Snow, 1960). Both sexes provide parental care, in the form of incubation and provisioning until fledging of chicks at approximately 55 days (Daunt *et al.*, 2007; Snow, 1960). During the non-breeding season, European shags remain coastal but may migrate variable distances from their breeding location. For example, on the Isle of May in Scotland, approximately 50% of the population migrates during the winter, with half the population remaining resident at the breeding area and half migrating up and down the entire East coast of the UK and, more rarely, across the North Sea to the Netherlands (Acker *et al.*, 2023).

Gulosus aristotelis has a range that covers most of the coastline of Europe, with high concentrations the Atlantic coast and Mediterranean, and smaller populations at its southern range limit on the coast of North Africa (GBIF Secretariat, 2023). With a European breeding population estimated at 152,000 (BirdLife International, 2021), European shags are listed as ‘least concern’ but decreasing on the IUCN red list (BirdLife International, 2018), with threats including climate change and extreme weather events, changing prey availability, pollution and disease (BirdLife International, 2018; Hicks *et al.*, 2019). Similarly to other seabirds, European shags are ecologically important, not only in their role as marine predators but as key indicators of marine ecosystem health and environmental change (Parsons *et al.*, 2008; Piatt & Sydeman, 2007).

Long term monitoring programmes of seabird colonies, including European shags, such as that carried out on the Isle of May in Scotland (from which this specimen originates), are crucial in building a picture of the interacting processes that

threaten seabirds. Therefore, the availability of this complete reference genome is a vital step in combining molecular tools with existing large life history datasets. For example, this genome will immediately aid the completion of an epigenetic clock, widening access to age data for birds that have not been individually marked as chicks, as well as the impact of stressors on biological aging where birds are of known age. This includes responses to infection and the genome sequence will be used to explore the molecular mechanisms of resistance and immunity against parasitism and disease, including long term chronic infections with parasites that are ubiquitous across individuals (Granroth-Wilding *et al.*, 2017), and species differences in response to recent Avian Influenza outbreaks.

The genome of the European shag was sequenced as part of the Darwin Tree of Life Project and the Vertebrate Genomes Project (VGP). Here we present a chromosomally complete genome sequence for *G. aristotelis*, based on one female specimen from the Isle of May National Nature Reserve, Scotland.

Genome sequence report

The genome was sequenced from a blood sample taken from one female *G. aristotelis* temporarily caught under licence from Isle of May National Nature Reserve, Scotland, UK (56.19, -2.57). A total of 42-fold coverage in Pacific Biosciences single-molecule HiFi long reads was generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 52 missing joins or mis-joins, reducing the scaffold number by 9.49%, and decreasing the scaffold N50 by 6.93%.

The final assembly has a total length of 1,279.1 Mb in 352 sequence scaffolds with a scaffold N50 of 78.9 Mb (Table 1). The snail plot in Figure 2 provides a summary of the assembly statistics, while the distribution of assembly scaffolds on GC proportion and coverage is shown in Figure 3. The cumulative assembly plot in Figure 4 shows curves for subsets of scaffolds assigned to different phyla. Most (96.21%) of the assembly sequence was assigned to 36 chromosomal-level scaffolds, representing 34 autosomes and the Z and W sex chromosomes. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 2). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The combined primary and alternate assemblies achieve an estimated QV of 62.1. The *k*-mer completeness is 97.25% for the primary assembly, 82.08% for the alternate haplotype, and 99.61% for the combined assemblies. The assembly has a BUSCO v5.3.2 completeness of 97.2% (single = 96.6%, duplicated = 0.6%), using the aves_odb10 reference set (*n* = 8,338). Based on the Earth BioGenome Project Report on Assembly Standards (September 2024), the EBP metric for the primary assembly is **7.C.Q61**, meeting the recommended reference standard.

Metadata for specimens, barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at <https://links.tol.sanger.ac.uk/species/126867>.

Genome annotation report

The *Gulosus aristotelis* genome assembly (GCA_949628215.1) was annotated at the European Bioinformatics Institute (EBI) using the Ensembl rapid annotation pipeline. The resulting annotation includes 26,595 transcribed mRNAs from 16,474 protein-coding and 1,001 non-coding genes (Table 1; https://rapid.ensembl.org/Phalacrocorax_aristotelis_GCA_949628215.1/Info/Index).

Methods

Sample acquisition and nucleic acid extraction

A blood sample was taken from a female *Gulosus aristotelis* (specimen ID SAN0001768, ToLID bGulAri2) from the Isle of May National Nature Reserve, Scotland, UK (latitude 56.19, longitude -2.57) on 2021-06-29. The species was identified by Hannah Ravenswater (University of Edinburgh), Fiona Greco (University of Edinburgh) and Sarah Burthe (UK Centre for Ecology & Hydrology). Capture occurred via crook at the nest site during late chick rearing, and the individual released to the same location. Blood sampling was conducted by appropriately trained personal license holders, acting under a UK Home Office Project License in accordance with the Animals (Scientific Procedures) Act 1986. The blood sample was collected via brachial venepuncture of the live specimen using a 25-gauge needle. 75 µl blood was placed in 500 µl 95% ethanol, and then frozen at -80 °C within 120 mins.

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) includes a sequence of core procedures: sample preparation; sample homogenisation, DNA extraction, fragmentation, and clean-up. In sample preparation, the bGulAri2 sample was weighed and dissected on dry ice (Jay *et al.*, 2023). The blood sample was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a).

HMW DNA was extracted using the Nanobind whole blood protocol (Pacific Biosciences *et al.*, 2023). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation (Oatley *et al.*, 2023): in brief, the method employs a 1.8X ratio of AMPure PB beads to sample to eliminate shorter fragments and concentrate the DNA. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from a blood sample from bGulAri2 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana protocol (do Amaral *et al.*, 2023). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

Table 1. Genome data for *Phalacrocorax aristotelis*, bGulAri2.1.

Project accession data		
Assembly identifier	bGulAri2.1	
Species	<i>Gulosus aristotelis</i> (previously <i>Phalacrocorax aristotelis</i>)	
Specimen	bGulAri2	
NCBI taxonomy ID	126867	
BioProject	PRJEB57282	
BioSample ID	SAMEA10059652	
Isolate information	bGulAri2, female: blood sample (DNA, Hi-C and RNA sequencing)	
Assembly metrics*		Benchmark
Consensus quality (QV)	Primary: 61.7; alternate: 62.3; combined: 62.1	≥ 40
k-mer completeness	Primary: 97.25%; alternate: 82.08%; combined: 99.61%	≥ 95%
BUSCO**	C:97.2%[S:96.6%,D:0.6%],F:0.5%,M:2.2%,n:8,338	<i>S</i> > 90%; <i>D</i> < 5%
Percentage of assembly mapped to chromosomes	96.21%	≥ 90%
Sex chromosomes	ZW	<i>localised homologous pairs</i>
Organelles	Mitochondrial genome: 18.61 kb	<i>complete single alleles</i>
Raw data accessions		
PacificBiosciences SEQUEL II	ERR10462081, ERR10462082	
Hi-C Illumina	ERR10466815	
PolyA RNA-Seq Illumina	ERR11606292	
Genome assembly		
Assembly accession	GCA_949628215.1	
<i>Accession of alternate haplotype</i>	GCA_949628205.1	
Span (Mb)	1,279.1	
Number of contigs	575	
Contig N50 length (Mb)	11.4	
Number of scaffolds	352	
Scaffold N50 length (Mb)	78.9	
Longest scaffold (Mb)	219.24	
Genome annotation		
Number of protein-coding genes	16,474	
Number of non-coding genes	1,001	
Number of gene transcripts	26,595	

* Assembly metric benchmarks are adapted from column VGP-2020 of "Table 1: Proposed standards and metrics for defining genome assembly quality" from Rhie *et al.* (2021).

** BUSCO scores based on the aves_odb10 BUSCO set using version 5.3.2. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/bGulAri2_1/dataset/bGulAri2_1/busco.

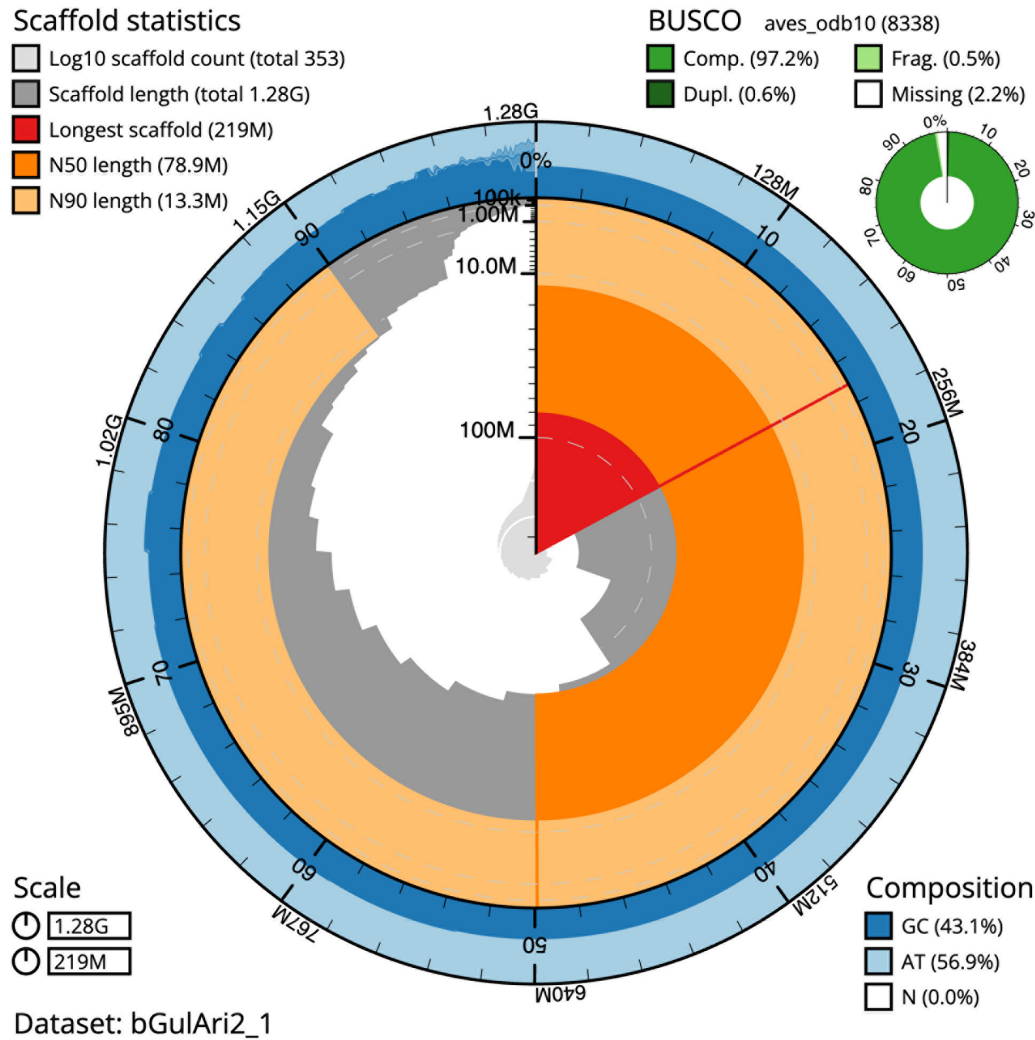


Figure 2. Genome assembly of *Gulus aristotelis*, (previously *Phalacrocorax aristotelis*), bGulAri2.1: metrics. The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 bins around the circumference with each bin representing 0.1% of the 1,279,134,750 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (219,240,020 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (78,889,319 and 13,321,506 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The outermost track of the plot shows the combined distribution of GC, AT and N percentages in the same bins as the inner plot, according to the key. A summary of complete, fragmented, duplicated and missing BUSCO genes in the aves_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bGulAri2_1/dataset/bGulAri2_1/snail.

Protocols developed by the WSI Tree of Life laboratory are publicly available on protocols.io (Denton *et al.*, 2023b).

Sequencing

Pacific Biosciences HiFi circular consensus DNA sequencing libraries were constructed according to the manufacturers' instructions. Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit. DNA and RNA sequencing was performed by the Scientific Operations core at the WSI on Pacific Biosciences SEQUEL II (HiFi) and Illumina NovaSeq 6000 (RNA-Seq) instruments. Hi-C data were also

generated from a blood sample from bGulAri2 using the Arima2 kit and sequenced on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

Assembly was carried out with Hifiasm (Cheng *et al.*, 2021) and haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020). The assembly was then scaffolded with Hi-C data (Rao *et al.*, 2014) using YaHS (Zhou *et al.*, 2023). The assembly was checked for contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as

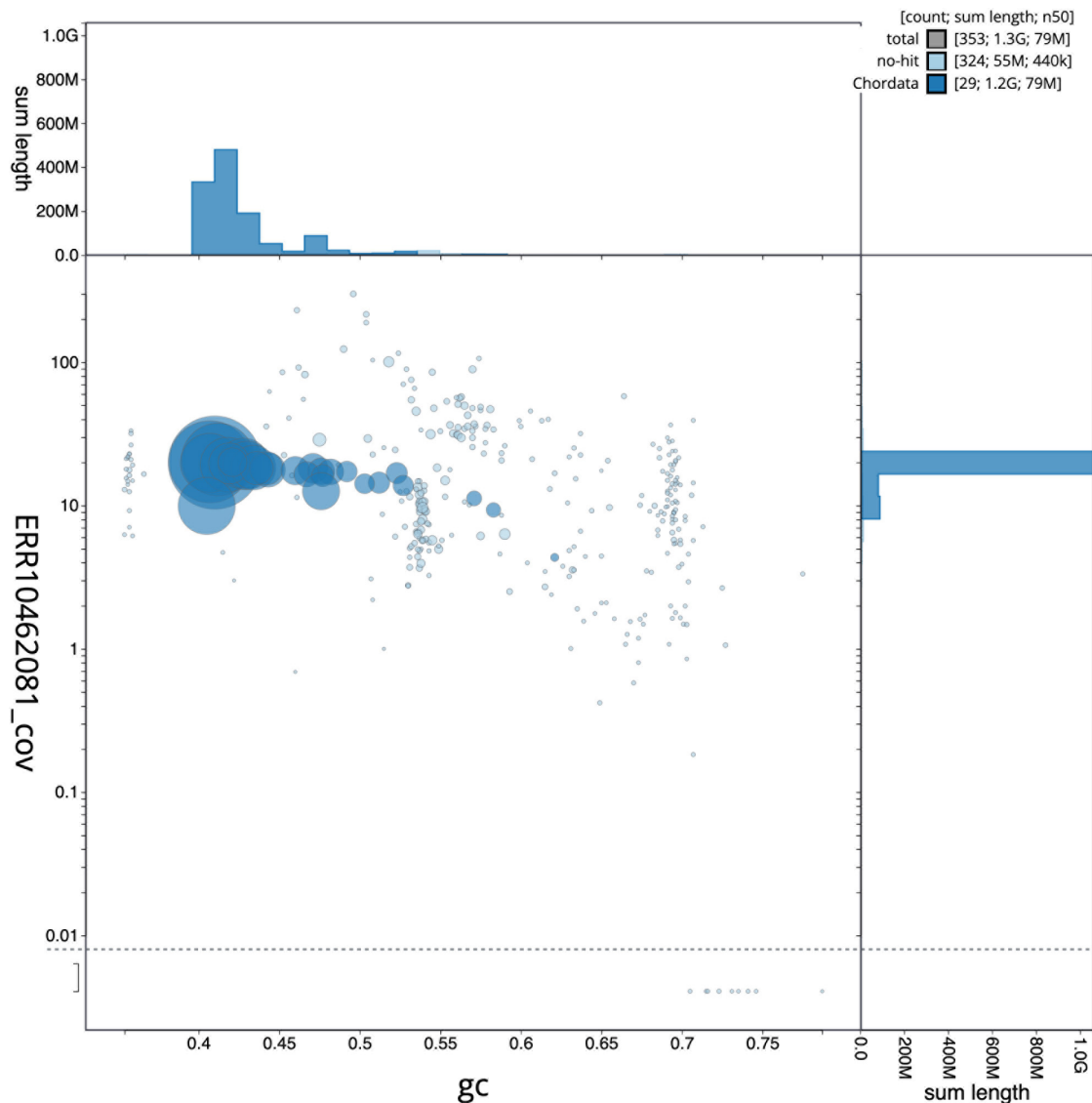


Figure 3. Genome assembly of *Gulosus aristotelis*, bGulAri2.1: BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bGulAri2_1/dataset/bGulAri2_1/blob.

described previously (Howe *et al.*, 2021). Manual curation was performed using gEVAL, HiGlass (Kerpedjiev *et al.*, 2018) and PretextView (Harry, 2022). The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were

calculated in Merqury.FK (Rhie *et al.*, 2020). The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021; Simão *et al.*, 2015) were calculated.

Table 3 contains a list of relevant software tool versions and sources.

Genome annotation

The Ensembl Genebuild annotation system (Aken *et al.*, 2016) at the EBI was used to generate annotation for the *Gulosus aristotelis* assembly (GCA_949628215.1). Annotation was created

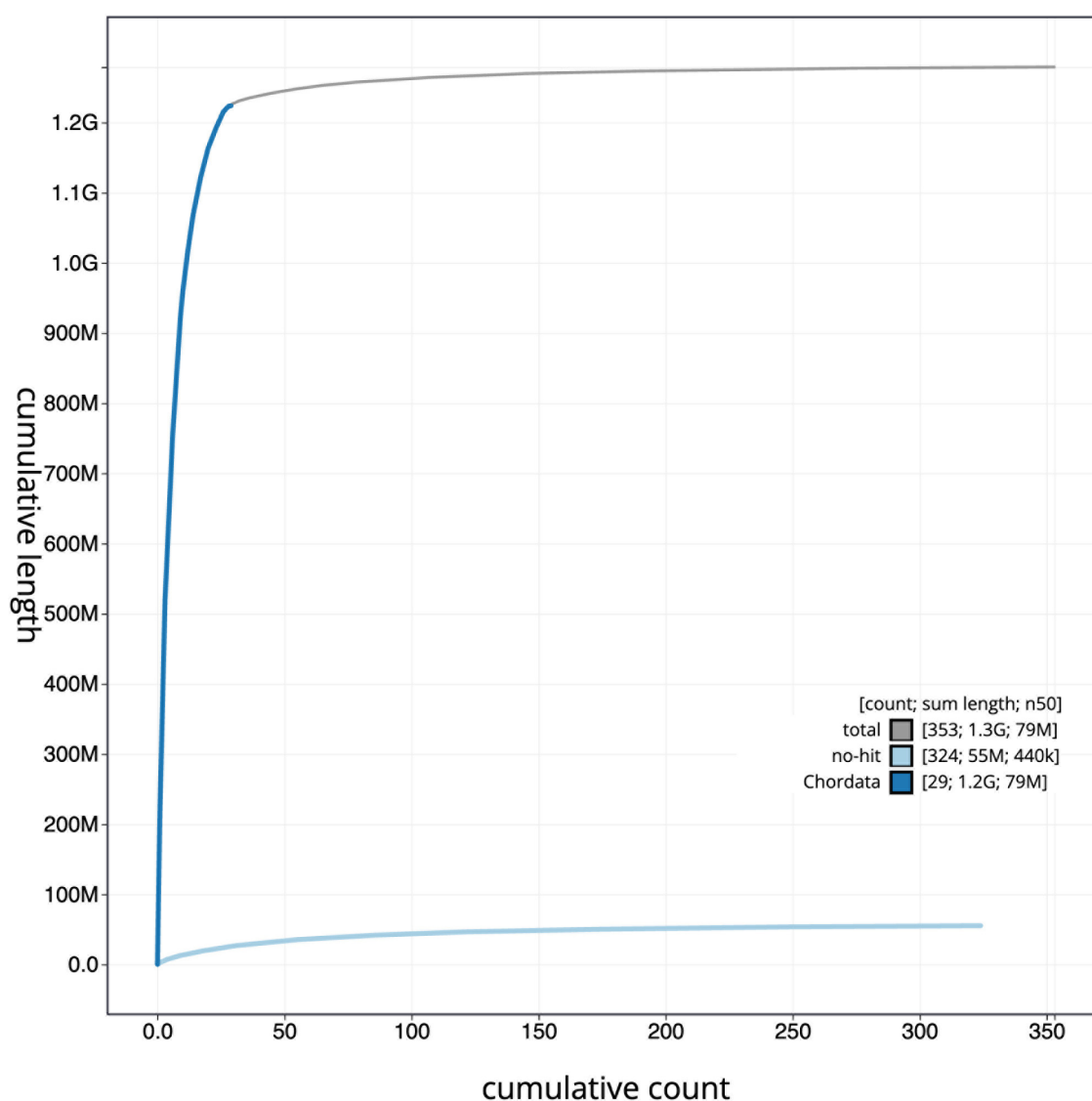


Figure 4. Genome assembly of *Gulosus aristotelis*, bGulAri2.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bGulAri2_1/dataset/bGulAri2_1/cumulative.

primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the ‘**Darwin Tree of Life Project Sampling Code of Practice**’, which can be found in full on the Darwin Tree of Life website [here](#). By agreeing with and signing up to the

Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project.

Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in

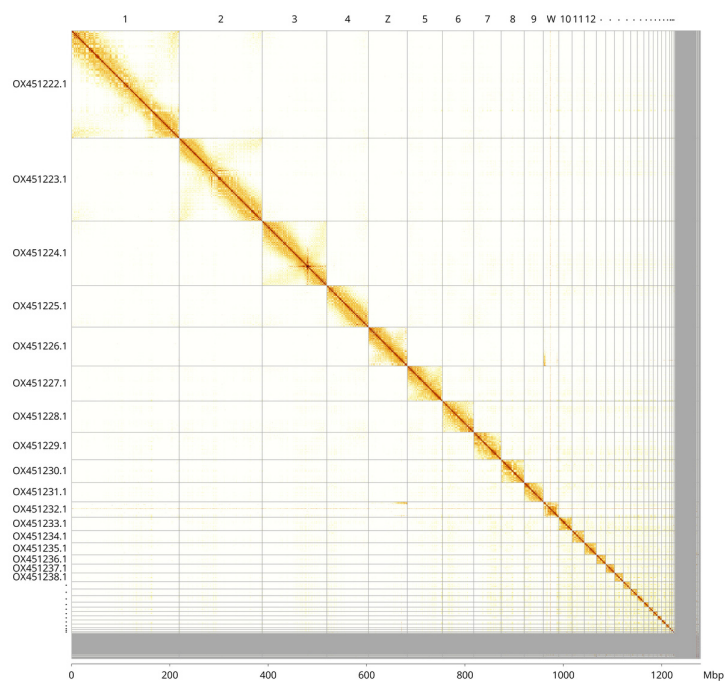


Figure 5. Genome assembly of *Gulosus aristotelis*, bGulAri2.1: Hi-C contact map of the bGulAri2.1 assembly, visualised using PretextView. Chromosomes are shown in order of size from left to right and top to bottom and are labelled along the axes with a megabase scale below. An interactive version of this figure in HiGlass may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/I/?d=AXw7sVdmRrWp1DFeFKZT-Q>.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Gulosus aristotelis*, bGulAri2.

INSDC accession	Chromosome	Length (Mb)	GC%
OX451222.1	1	219.24	41.0
OX451223.1	2	168.56	40.5
OX451224.1	3	131.34	41.0
OX451225.1	4	84.76	40.5
OX451227.1	5	71.72	42.0
OX451228.1	6	63.33	43.0
OX451229.1	7	55.81	43.0
OX451230.1	8	46.83	43.5
OX451231.1	9	39.05	42.0
OX451233.1	10	26.91	44.5
OX451234.1	11	24.85	44.0
OX451235.1	12	24.68	43.5
OX451236.1	13	19.06	47.0
OX451237.1	14	17.93	42.0
OX451238.1	15	17.37	46.0
OX451239.1	16	14.99	47.5
OX451240.1	17	13.73	48.0

INSDC accession	Chromosome	Length (Mb)	GC%
OX451241.1	18	13.32	46.5
OX451242.1	19	9.84	47.5
OX451243.1	20	9.08	51.0
OX451244.1	21	8.89	52.5
OX451245.1	22	8.36	49.0
OX451246.1	23	8.19	52.5
OX451247.1	24	7.62	50.5
OX451248.1	25	3.89	57.0
OX451249.1	26	3.57	58.5
OX451250.1	27	2.68	47.5
OX451251.1	28	1.64	59.0
OX451252.1	29	1.11	54.5
OX451253.1	30	0.71	62.0
OX451254.1	31	0.63	57.5
OX451255.1	32	0.33	63.0
OX451256.1	33	0.23	61.5
OX451257.1	34	0.23	59.5
OX451232.1	W	31.35	47.5
OX451226.1	Z	78.89	40.5
OX451258.1	MT	0.02	44.5

Table 3. Software tools: versions and sources.

Software tool	Version	Source
BlobToolKit	4.1.7	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.3.2	https://gitlab.com/ezlab/busco
gEVAL	-	https://geval.org.uk/
Hifiasm	0.16.1	https://github.com/chhylp123/hifiasm
HiGlass	1.11.6	https://github.com/higlass/higlass
Mercury.FK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
PretextView	0.2	https://github.com/wtsi-hpag/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
YaHS	yahs-1.1.91eebc2	https://github.com/c-zhou/yahs

doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: *Gulosus aristotelis* (European shag). Accession number PRJEB57282; <https://identifiers.org/ena.embl/PRJEB57282> (Wellcome Sanger Institute, 2023). The genome sequence is released openly for reuse. The *Gulosus aristotelis* genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project and the Vertebrate Genomes Project (VGP). All raw sequence data and the

assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1.

Author information

Members of the Darwin Tree of Life Barcoding collective are listed here: <https://doi.org/10.5281/zenodo.4893703>.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.10066175>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.10043364>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.10066637>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.5013541>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>.

References

Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Acker P, Daunt F, Wanless S, et al.: **Additive genetic and environmental variation interact to shape the dynamics of seasonal migration in a wild bird population.** *Evolution.* 2023; **77**(10): 2128–2143.
[PubMed Abstract](#) | [Publisher Full Text](#)

Aebischer NJ: **Retrospective investigation of an ecological disaster in the shag, *Phalacrocorax aristotelis*: a general method based on long-term marking.** *J Anim Ecol.* 1986; **55**(2): 613–629.
[Publisher Full Text](#)

Aken BL, Ayling S, Barrell D, et al.: **The ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

- Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor® 3 for LI PacBio.** *protocols.io.* 2023. [Publisher Full Text](#)
- BirdLife International: **Gulosus aristotelis.** The IUCN Red List of Threatened Species. 2018.
- BirdLife International: **European Red List of Birds.** Luxembourg: Office for Official Publications of the European Communities, 2021. [Reference Source](#)
- Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL—a web-based browser for evaluating genome assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–2510. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Croxall JP: **Seabirds: feeding ecology and role in marine ecosystems.** Cambridge University Press, 1987; 8. [Reference Source](#)
- Daunt F, Afanasyev V, Adam A, *et al.*: **From cradle to early grave: juvenile mortality in European shags *Phalacrocorax aristotelis* results from inadequate development of foraging proficiency.** *Biol Lett.* 2007; **3**(4): 371–374. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Daunt F, Monaghan P, Wanless S, *et al.*: **Sons and daughters: age-specific differences in parental rearing capacities.** *Funct Ecol.* 2001; **15**(2): 211–216. [Publisher Full Text](#)
- Denton A, Oatley G, Cornwell C, *et al.*: **Sanger Tree of Life Sample homogenisation: PowerMash.** *protocols.io.* 2023a. [Publisher Full Text](#)
- Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b. [Publisher Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows.** *Nat Biotechnol.* 2017; **35**(4): 316–319. [PubMed Abstract](#) | [Publisher Full Text](#)
- do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *protocols.io.* 2023. [Publisher Full Text](#)
- GBIF Secretariat: ***Phalacrocorax aristotelis* (Linnaeus, 1761).** *GBIF Backbone Taxonomy.* 2023; [Accessed 15 February 2024]. [Reference Source](#)
- Granroth-Wilding HMV, Burthe SJ, Lewis S, *et al.*: **Parasitism in early life: environmental conditions shape within-brood variation in responses to infection.** *Ecol Evol.* 2014; **4**(17): 3408–3419. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Granroth-Wilding HMV, Daunt F, Cunningham EJA, *et al.*: **Between-individual variation in nematode burden among juveniles in a wild host.** *Parasitology.* 2017; **144**(2): 248–258. [PubMed Abstract](#) | [Publisher Full Text](#)
- Grémillet D, Tuschy I, Kierspel M: **Body temperature and insulation in diving Great Cormorants and European Shags.** *Funct Ecol.* 1998; **12**(3): 386–394. [Publisher Full Text](#)
- Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hall ME: **Senescence and reproductive performance in the European shag (*Phalacrocorax aristotelis*).** University of Glasgow (United Kingdom), 2004. [Reference Source](#)
- Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps.** 2022; [Accessed 19 October 2022]. [Reference Source](#)
- Herborn KA, Heidinger BJ, Boner W, *et al.*: **Stress exposure in early post-natal life reduces telomere length: an experimental demonstration in a long-lived seabird.** *Proc Biol Sci.* 2014; **281**(1782): 20133151. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hicks O, Green JA, Daunt F, *et al.*: **Sublethal effects of natural parasitism act through maternal, but not paternal, reproductive success in a wild population.** *Ecology.* 2019; **100**(8): e02772. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* Oxford University Press, 2021; **10**(1): g1aa153. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Howells R, Burthe S, Green J, *et al.*: **From days to decades: short- and long-term variation in environmental conditions affect offspring diet composition of a marine top predator.** *Mar Ecol Prog Ser.* 2017; **583**: 227–242. [Publisher Full Text](#)
- Howells RJ, Burthe SJ, Green JA, *et al.*: **Pronounced long-term trends in year-round diet composition of the European shag *Phalacrocorax aristotelis*.** *Mar Biol.* 2018; **165**(12): 188. [Publisher Full Text](#)
- Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: **Sanger Tree of Life Sample preparation: triage and dissection.** *protocols.io.* 2023. [Publisher Full Text](#)
- Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Manni M, Berkeley MR, Seppay M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Oatley G, Sampaio F, Howard C: **Sanger Tree of Life fragmented DNA clean up: automated SPRI V.1.** *protocols.io.* 2023. [Publisher Full Text](#)
- Pacific Biosciences, Denton A, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA extraction: manual nucleated blood Nanobind®.** *protocols.io.* 2023; [Accessed 5 January 2024]. [Publisher Full Text](#)
- Parsons M, Mitchell I, Butler A, *et al.*: **Seabirds as indicators of the marine environment.** *ICES J Mar Sci.* 2008; **65**(8): 1520–1526. [Publisher Full Text](#)
- Piatt I, Sydeman W: **Seabirds as indicators of marine ecosystems.** *Mar Ecol Prog Ser.* 2007; **352**: 199–204. [Publisher Full Text](#)
- Rao SSP, Huntley MH, Durd NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rhie A, Walenz BP, Koren S, *et al.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212. [PubMed Abstract](#) | [Publisher Full Text](#)
- Snow B: **The breeding biology of the Shag *Phala crocorax aristotelis* on the Island of Lundy, Bristol Channel.** *Ibis.* 1960; **102**(4): 554–575. [Publisher Full Text](#)
- Thanou E, Giokas S, Goutner V, *et al.*: **Efficiency and accuracy of PCR-Based sex determination Methods in the European phalacrocoracidae.** *Ann Zool Fenn.* 2013; **50**(1–2): 52–63. [Publisher Full Text](#)
- Thanou E, Sponza S, Nelson EJ, *et al.*: **Genetic structure in the European endemic seabird, *Phalacrocorax aristotelis*, shaped by a complex interaction of historical and contemporary, physical and nonphysical drivers.** *Mol Ecol.* 2017; **26**(10): 2796–2811. [PubMed Abstract](#) | [Publisher Full Text](#)
- Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Vasimuddin M, Misra S, Li H, *et al.*: **Efficient Architecture-Aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324. [Publisher Full Text](#)
- Velando A, Freire J: **Intercolony and seasonal differences in the breeding diet of European shags on the Galician coast (NW Spain).** *Mar Ecol Prog Ser.* 1999; **188**: 225–236. [Publisher Full Text](#)
- Velando A, Freire J: **Can the central-periphery distribution become general in seabird colonies? Nest spatial pattern in the European Shag.** *Condor.* 2001; **103**: 544–554.
- Wanless S, Harris MP: ***Phalacrocorax aristotelis* Shag.** Birds of the Western Palearctic Update. 1997; **1**(1): 3–13. [Reference Source](#)
- Wellcome Sanger Institute: **The genome sequence of the European shag, *Gulosus aristotelis* (previously *Phalacrocorax aristotelis*) (Linnaeus, 1761).** European Nucleotide Archive. [dataset], accession number PRJEB57282, 2023.
- Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* Edited by C. Alkan, 2023; **39**(1): btac808. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 16 July 2024

<https://doi.org/10.21956/wellcomeopenres.23361.r86201>

© 2024 Kan X. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Xianzhao Kan

¹ Anhui Normal University, Wuhu, Anhui, China

² Anhui Normal University, Wuhu, Anhui, China

The manuscript reports a genome assembly from an individual female, previously known as *Phalacrocorax aristotelis*, (the European shag; Chordata; Aves; Pelecaniformes; Phalacrocoracidae). The availability of this complete reference genome is a vital step in combining molecular tools with existing large life history datasets. Therefore, I think this is very meaningful and suitable for indexing in Wellcome Open Research. Some of my specific comments are below.

Specific comments:

1. In the second paragraph of Background: "European Shags" should be "European shags".
2. In the fourth paragraph of Background: "*G.aristotelis*" should be "*G. aristotelis*".
3. In the first paragraph of Genome sequence report section: "The genome was sequenced from a blood sample taken from one female *Gulosus aritotelis* (previously *Phalacrocorax aristotelis*)...", *Gulosus aritotelis* can be abbreviated to "*G. aristotelis*", and "previously *Phalacrocorax aristotelis*" can be removed as it is already covered in the Background section.
4. In Table 1, the size of mitochondrial genome in 18.61 kb, which is longer than typical avian mitogenomes. Please analyze the possible reasons for this.
5. In Table 2, the authors should further investigate the correlation between the chromosome length and GC content percentage, and analyze the possible reasons. Specifically, they should explain why the five chromosomes shorter than 1 Mb have a very high GC content percentage.
6. The figure legend for Figure 1, "Photograph of a *Gulosus aristotelis* (previously *Phalacrocorax aristotelis*) (bGulAri2) from same study..." should be changed to "Photograph of a *Gulosus aristotelis* (previously *Phalacrocorax aristotelis*) (bGulAri2) from the same study..."

7. The figure legend for Figure 2 should be modified to "Figure 2. Genome assembly of *Gulosus aristotelis* (previously *Phalacrocorax aristotelis*), bGulAri2.1: metrics." Furthermore, "The blue and pale-blue area around the outside of the plot shows ..." had better changed into "The blue and pale-blue areas around the outside of the plot show ..."
8. In the first paragraph of Sample acquisition and nucleic acid extraction section: "The species was identified by Hannah Ravenswater (University of Edinburgh) and, Fiona Greco (University of Edinburgh) and Sarah Burthe (UK Centre for Ecology & Hydrology)." Should be "The species was identified by Hannah Ravenswater (University of Edinburgh), Fiona Greco (University of Edinburgh), and Sarah Burthe (UK Centre for Ecology & Hydrology)."

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics, Bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 03 July 2024

<https://doi.org/10.21956/wellcomeopenres.23361.r86203>

© 2024 Rawlence N. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Nicolas J. Rawlence

¹ University of Otago, Dunedin, New Zealand

² University of Otago, Dunedin, New Zealand

Ravenswater *et al.* have sequenced the nuclear and mitochondrial genome of the European shag *Gulosus aristotelis*, which will be a useful genomic resource for studying the evolution of

cormorants and shags, including flight loss within the group. I hope my comments are helpful in revising this data note.

The mitogenome that was produced is 18.61 kb in length. This suggests that there is a duplication within the mitogenome, probably of the Control Region/D-loop and surrounding genes (e.g. ND5, Cyt b) and tRNAs. It would be good to briefly comment on the increased size of the mitogenome and the reasons for it.

The whole manuscript needs a thorough proof read as evidenced by the revisions noted below:

1. "Megabases in span". Can you please change this to "megabases in size".
2. Please make genus and species names in italics.
3. "European shags have dark-green plumage in adulthood but that can often appear black, a yellow gape, green eyes, black legs and feet and a crest during the breeding season (Wanless & Harris, 1997)" should be "European shags have dark-green plumage in adulthood but can often appear black, a yellow gape, green eyes, black legs and feet, and a crest during the breeding season (Wanless & Harris, 1997)."
4. Please be consistent in use of "European Shag" vs "European shag".
5. "both temporal and spatial variability consumption of prey species". Something is missing here and needs rewording.
6. "an important seabird breeding colony where shags have been intensively monitored since the 1970s". Put this section in parentheses, otherwise it is a very long sentence.
7. "For example, on the Isle of May in Scotland, approximately 50% of the population migrates during the winter, with half the population remaining resident at the breeding area and half migrating up and down the entire East coast of the UK and, more rarely, across the North Sea to the Netherlands (Acker et al., 2023)" I would mention Scotland in the first instance of Isle of May, not here or in other places latter in the text. "East coast" does not need a capital E.
8. "*G. aristotelis* has a range that covers". This is the start of a paragraph, so the genus species name needs to be spelled out in full.
9. "with high concentrations the Atlantic coast and Mediterranean". There are some words missing here, please insert them.
10. "The genome was sequenced from a blood sample taken from one female *Gulosus aritotelis* (previously *Phalacrocorax aristotelis*)" Please make sure to spell the genus and species name correctly. There are many instances where this is not the case.
11. The figure three caption has the wrong genus.
12. "The species was identified by Hannah Ravenswater (University of Edinburgh) and, Fiona Greco (University of Edinburgh) and Sarah Burthe (UK Centre for Ecology & Hydrology)". Please fix the grammar in this sentence.

Is the rationale for creating the dataset(s) clearly described?

Yes

Are the protocols appropriate and is the work technically sound?

Yes

Are sufficient details of methods and materials provided to allow replication by others?

Yes

Are the datasets clearly presented in a useable and accessible format?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolutionary genetics of cormorants and shags

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.
