DATA NOTE

# The genome sequence of the Common Goldeneye, *Bucephala clangula* (Linnaeus, 1758)

[version 1; peer review: 2 approved]

Rosa Lopez Colom[1], Michelle F. O'Brien[1],
Natural History Museum Genome Acquisition Lab,
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

[1]Wildfowl and Wetlands Trust, Slimbridge, England, UK

**v1** **First published:** 20 May 2025, **10**:258
https://doi.org/10.12688/wellcomeopenres.24275.1
**Latest published:** 20 May 2025, **10**:258
https://doi.org/10.12688/wellcomeopenres.24275.1

## Abstract

We present a genome assembly from a male *Bucephala clangula* (Common Goldeneye; Chordata; Aves; Anseriformes; Anatidae). The genome sequence has a total length of 1,190.92 megabases. Most of the assembly (95.05%) is scaffolded into 40 chromosomal pseudomolecules, including the Z sex chromosome. The mitochondrial genome has also been assembled, with a length of 16.63 kilobases.

## Keywords

Bucephala clangula, Common Goldeneye, genome sequence, chromosomal, Anseriformes

This article is included in the Tree of Life gateway.

**Open Peer Review**

**Approval Status** ✓ ✓

| | 1 | 2 |
|---|---|---|
| **version 1** 20 May 2025 | ✓ view | ✓ view |

1. **Rauri C.K. Bowie**, University of California Berkeley, Berkeley, USA
2. **Martin Pippel**, Uppsala University, Husargatan 3, Uppsala, Sweden

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: Lopez Colom R**: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **O'Brien MF**: Investigation, Resources, Writing – Original Draft Preparation, 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 (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328 ].
*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*



**How to cite this article:** Lopez Colom R, O'Brien MF, Natural History Museum Genome Acquisition Lab *et al.* **The genome sequence of the Common Goldeneye, *Bucephala clangula* (Linnaeus, 1758) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:258 https://doi.org/10.12688/wellcomeopenres.24275.1

**First published:** 20 May 2025, **10**:258 https://doi.org/10.12688/wellcomeopenres.24275.1

## 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; Galloanserae; Anseriformes; Anatidae; Anatinae; *Bucephala*; *Bucephala clangula* (Linnaeus, 1758) (NCBI: txid107022)

## Background

The Common Goldeneye (*Bucephala clangula*) is a medium-sized duck found across the Northern Hemisphere. In the UK, Goldeneyes are primarily winter visitors, arriving in October from Scandinavia, with breeding populations limited to Scotland (Humble & McGill, 2019).

Although sexual dimorphism is present in this species, both males and females have distinct golden-coloured eyes and orange legs. The male's eyes stand out due to its dark green (almost black) head, white cheeks, and black-and-white striped wings. Female Goldeneyes have a more brownish hue to their head plumage, with a mottled grey-brown and white body (Humble & McGill, 2019). They are approximately 46 cm in length, have a wingspan of up to 80 cm, weigh between 650 g and 1,200 g, and can be observed in intertidal, freshwater and wetland habitats (RSPB, 2023). Although the maximum recorded age of a ringed bird is 12 years, on average, they have a lifespan of six years, with breeding typically beginning at two years of age (British Trust for Ornithology, 2015).

During the breeding season, from April onwards, courtship displays can be observed, with males throwing their heads back and kicking water into the air while producing a whistling sound (Humble & McGill, 2019). Goldeneyes are cavity-nesting waterfowl and therefore rely on old hollow trees or cavities created by other species (e.g. squirrels, woodpeckers, etc.). Due to this nesting limitation, a common conservation technique is to place wooden nest boxes to support breeding success (Pöysä & Pöysä, 2002). From late April to early June, eggs are laid, with the average clutch being from 9 to 11 eggs (Owen *et al.*, 1986). Within one to two days of hatching, ducklings leap from their tree cavity nests, which can be several metres high, to the ground, where they then follow their mother to bodies of water. This breeding behaviour is better developed in mature forests, where nesting sites are available and a well-developed understory is present. Observations of different types of nesting parasitism have been recorded, including egg-laying in the nests of other female goldeneyes in addition to the females' own nests, increasing the brood output (Åhlund & Andersson, 2001).

Apart from being cavity-nesting waterfowl, they are also part of the diving duck family and can forage depths of 7 m for aquatic invertebrates, molluscs, crustaceans, as well as small fish (Owen *et al.*, 1986).

Although the Common Goldeneye has a conservation IUCN status of 'least concern' (BirdLife International, 2018) with wetland habitats increasingly impacted by environmental changes, studying the species' genetics could offer valuable insights for conservation efforts and help ensure its long-term survival.

Here we present a chromosomally complete genome sequence for *Bucephala clangula*, based on a specimen from Gloucester, England, United Kingdom (Figure 1).

## Genome sequence report

### Sequencing data

The genome of a specimen of *Bucephala clangula* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 30.20 Gb (gigabases) from 2.55 million reads, which were used to assemble the genome. GenomeScope analysis estimated the haploid genome size at 1,177.45 Mb, with a heterozygosity of 0.42% and repeat content of 11.98%. These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 45 coverage. Hi-C sequencing produced 54.51 Gb from 360.97 million reads, used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

### Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 56 misjoins and missing joins and removed two haplotypic duplications. These interventions decreased the scaffold count by 5.11%. The final assembly has a total length of


**Figure 1. Photograph of the *Bucephala clangula* (bBucCla1) specimen used for genome sequencing.**

**Table 1. Specimen and sequencing data for *Bucephala clangula*.**

| **Project information** | | | |
|---|---|---|---|
| **Study title** | Bucephala clangula (common goldeneye) | | |
| **Umbrella BioProject** | PRJEB68286 | | |
| **Species** | *Bucephala clangula* | | |
| **BioSpecimen** | SAMEA112468123 | | |
| **NCBI taxonomy ID** | 107022 | | |
| **Specimen information** | | | |
| **Technology** | **ToLID** | **BioSample accession** | **Organism part** |
| **PacBio long read sequencing** | bBucCla1 | SAMEA112468148 | muscle |
| **Hi-C sequencing** | bBucCla1 | SAMEA112468148 | muscle |
| **RNA sequencing** | bBucCla1 | SAMEA112468148 | muscle |
| **Sequencing information** | | | |
| **Platform** | **Run accession** | **Read count** | **Base count (Gb)** |
| **Hi-C Illumina NovaSeq 6000** | ERR12259841 | 3.61e+08 | 54.51 |
| **PacBio Sequel IIe** | ERR12257416 | 0.00e+00 | 0.0 |
| **PacBio Sequel IIe** | ERR12257417 | 2.55e+06 | 30.2 |
| **RNA Illumina NovaSeq 6000** | ERR12321241 | 7.77e+07 | 11.74 |

1,190.92 Mb in 537 scaffolds, with 454 gaps, and a scaffold N50 of 66.36 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (95.05%) was assigned to 40 chromosomal-level scaffolds, representing 39 autosomes and the Z sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

### Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 61.3. The *k*-mer completeness is 91.65% for the primary haplotype and 71.68% for the alternate haplotype; and 99.48% for the combined primary and alternate assemblies. BUSCO v.5.5.0 analysis using the aves_odb10 reference set ($n$ = 8,338) identified 97.0% of the expected gene set (single = 96.7%, duplicated = 0.3%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The primary assembly achieves the EBP reference standard of **6.C.Q60**.

## Methods

### Sample acquisition

Several small samples of pectoral muscle were collected from a deceased Common Goldeneye, *Bucephala clangula*, specimen ID NHMUK014561651 (ToLID bBucCla1). This diving duck species was collected from a private collection in Gloucestershire in 2022, and stored in –20 °C freezers prior to sampling. The specimen was a male.

**Table 2. Genome assembly data for *Bucephala clangula*.**

| **Genome assembly** | | |
|---|---|---|
| Assembly name | bBucCla1.1 | |
| Assembly accession | GCA_964059595.1 | |
| *Alternate haplotype accession* | *GCA_964059605.1* | |
| Assembly level for primary assembly | chromosome | |
| Span (Mb) | 1,190.92 | |
| Number of contigs | 991 | |
| Number of scaffolds | 537 | |
| Longest scaffold (Mb) | 130.64 | |
| **Assembly metric** | **Measure** | ***Benchmark*** |
| Contig N50 length | 4.27 Mb | *≥ 1 Mb* |
| Scaffold N50 length | 66.36 Mb | *= chromosome N50* |
| Consensus quality (QV) | Primary: 60.8; alternate: 61.6; combined: 61.3 | *≥ 40* |
| *k*-mer completeness | Primary: 91.65%; alternate: 71.68%; combined: 99.48% | *≥ 95%* |
| BUSCO* | C:97.0%[S:96.7%,D:0.3%], F:0.5%,M:2.5%,n:8,338 | *S > 90%; D < 5%* |
| Percentage of assembly assigned to chromosomes | 95.05% | *≥ 90%* |
| Sex chromosomes | Z | *localised homologous pairs* |
| Organelles | Mitochondrial genome: 16.63 kb | *complete single alleles* |

* BUSCO scores based on the aves_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

## Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification (Howard *et al.*, 2025). Detailed protocols are available on protocols.io (Denton *et al.*, 2023). The bBucCla1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the muscle was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023).

HMW DNA was extracted in the WSI Scientific Operations core using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023). The 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, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the extracted DNA had a Qubit concentration of 17.52 ng/μL and a yield of 823.44 ng. Spectrophotometric measurements indicated 260/280 and 260/230 ratios of 2.08 and 3.58, respectively.

RNA was extracted from muscle tissue of bBucCla1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol (do Amaral *et al.*, 2023). The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the

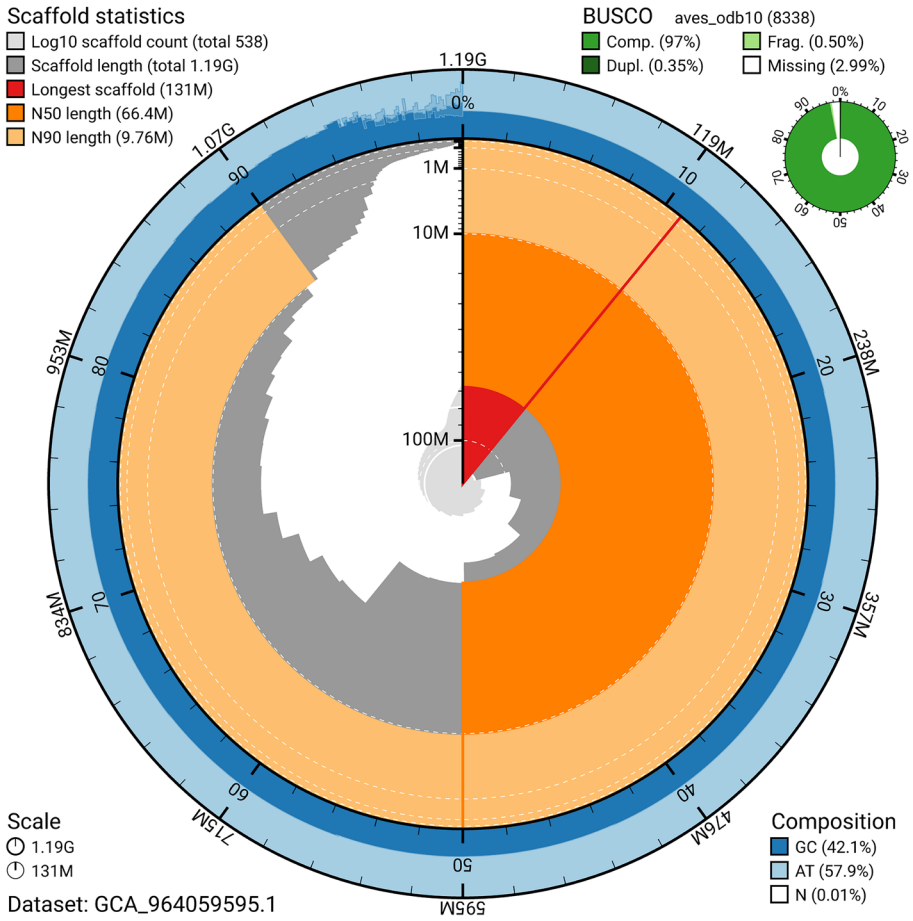


**Figure 2. Genome assembly of *Bucephala clangula*, bBucCla1.1: metrics.** The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the aves_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964059595.1/dataset/GCA_964059595.1/snail.

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.

## Hi-C sample preparation and crosslinking

Hi-C data were generated from the muscle of the bBucCla1 sample using the Arima-HiC v2 kit (Arima Genomics) with 20–50 mg of frozen tissue (stored at –80 °C). As per manufacturer's instructions, tissue was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration, and a final formaldehyde concentration of 2%. The tissue was then homogenised using the Diagnocine Power Masher-II. The crosslinked DNA was digested using a restriction enzyme master mix, then biotinylated and ligated. A clean up was performed with SPRIselect beads prior to library preparation. DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit, and sample biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

## Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

### *PacBio HiFi*

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low-input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Size-selection and clean-up were carried out using diluted AMPure

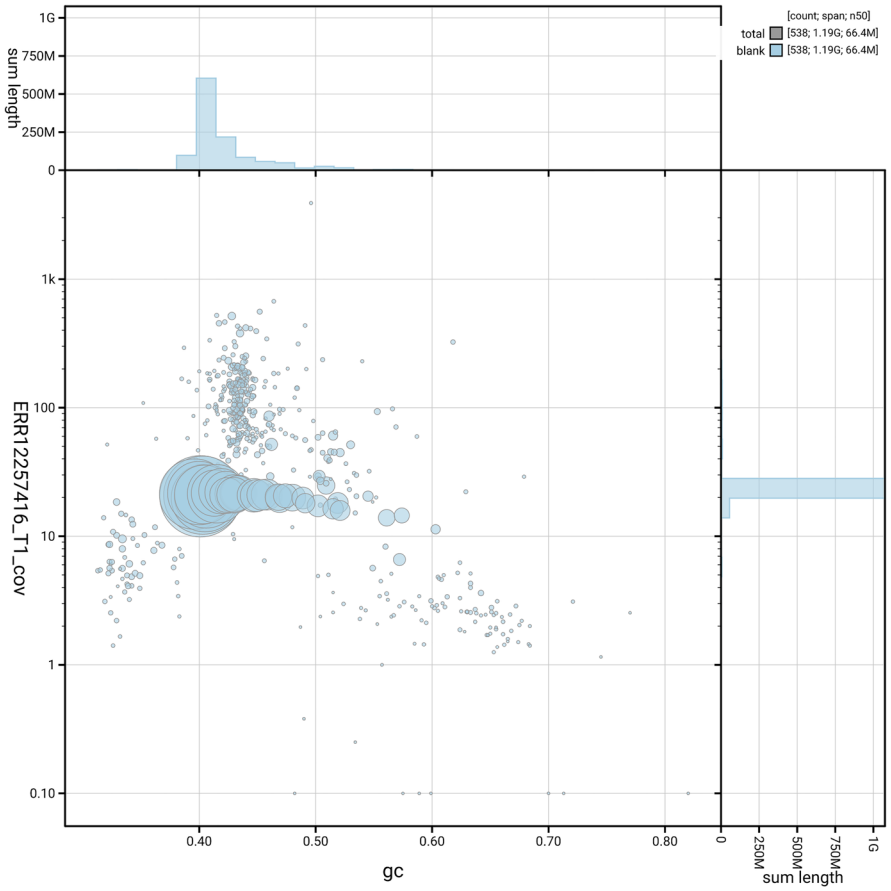


**Figure 3. Genome assembly of *Bucephala clangula*, bBucCla1.1: BlobToolKit GC-coverage plot.** Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964059595.1/dataset/GCA_964059595.1/blob.

PB beads (Pacific Biosciences). DNA concentration was quantified using the Qubit Fluorometer v4.0 (ThermoFisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and the gDNA 55kb BAC analysis kit.

Samples were sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, as well as perform primary and secondary analysis of the data upon completion.

***Hi-C***

For Hi-C library preparation, the biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size-selected to 400–600 bp using SPRISelect beads. DNA was then enriched using Arima-HiC v2 Enrichment beads. The NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) was used for end repair, A-tailing, and adapter ligation, following a modified protocol in which library preparation is carried out while the DNA remains bound to the enrichment beads. PCR amplification was performed using KAPA HiFi HotStart mix and custom dual-indexed adapters (Integrated DNA Technologies) in a 96-well plate format. Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, samples were amplified for 10–16 PCR cycles. Post-PCR clean-up was carried out using SPRISelect beads. The libraries were quantified using the Accuclear Ultra High Sensitivity dsDNA Standards Assay kit (Biotium) and normalised to 10 ng/μL before sequencing. Hi-C sequencing was performed on the Illumina NovaSeq 6000 instrument.

***RNA***

Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit, following the manufacturer's instructions. RNA sequencing was performed on the Illumina NovaSeq 6000 instrument.

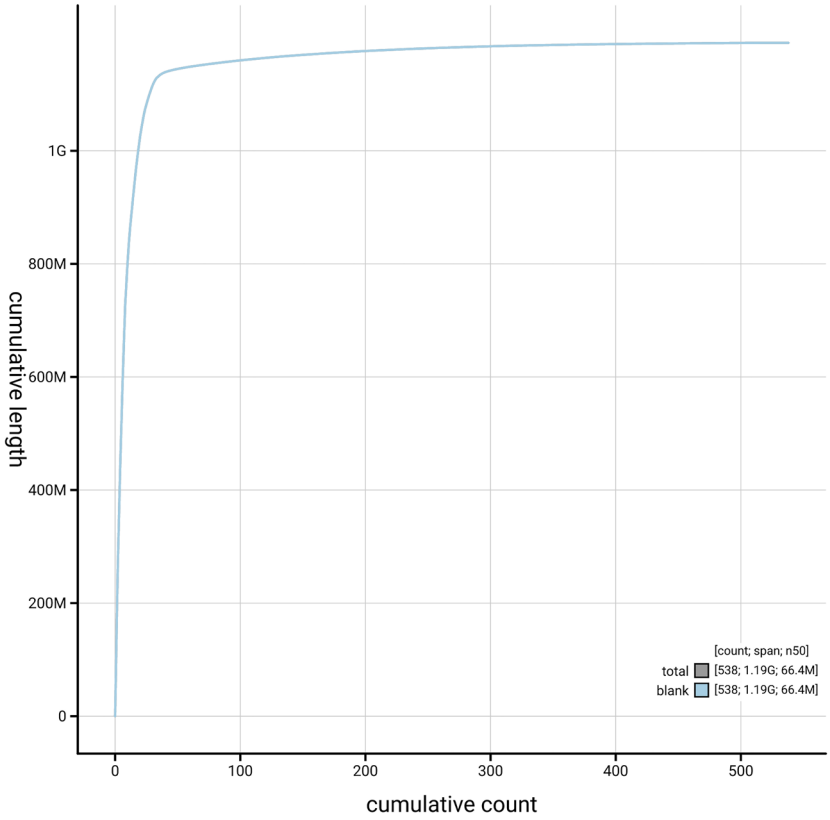


**Figure 4. Genome assembly of *Bucephala clangula*, bBucCla1.1: BlobToolKit cumulative sequence plot.** The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_964059595.1/dataset/GCA_964059595.1/cumulative.

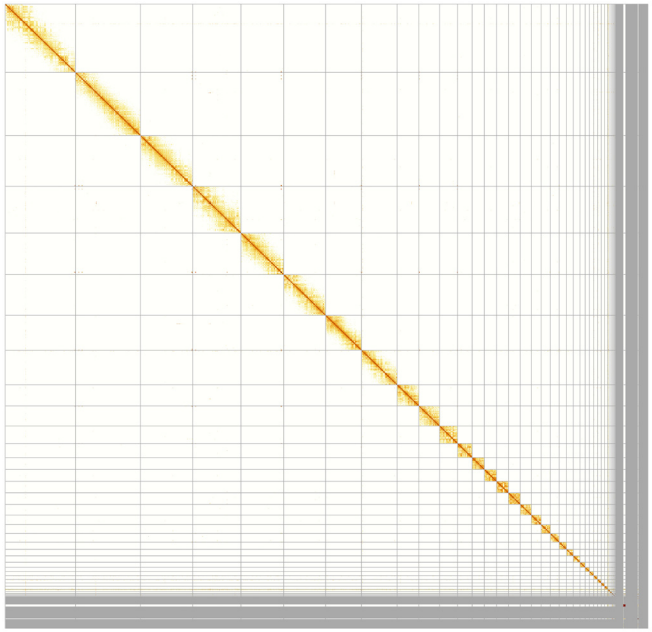

**Figure 5. Genome assembly of *Bucephala clangula*: Hi-C contact map of the bBucCla1.1 assembly, generated using PretextSnapshot.** Chromosomes are shown in order of size.

**Table 3. Chromosomal pseudomolecules in the genome assembly of *Bucephala clangula*, bBucCla1.**

| INSDC accession | Name | Length (Mb) | GC% |
|---|---|---|---|
| OZ060833.1 | 1 | 130.64 | 40 |
| OZ060834.1 | 2 | 120.4 | 40.5 |
| OZ060835.1 | 3 | 96.81 | 39.5 |
| OZ060837.1 | 4 | 79.0 | 40.5 |
| OZ060838.1 | 5 | 77.85 | 40 |
| OZ060839.1 | 6 | 66.36 | 40.5 |
| OZ060840.1 | 7 | 65.89 | 41.5 |
| OZ060841.1 | 8 | 40.16 | 41.5 |
| OZ060842.1 | 9 | 38.61 | 42 |
| OZ060843.1 | 10 | 33.13 | 42.5 |
| OZ060844.1 | 11 | 27.07 | 43 |
| OZ060845.1 | 12 | 22.62 | 43.5 |
| OZ060846.1 | 13 | 22.57 | 43.5 |
| OZ060847.1 | 14 | 22.03 | 42.5 |
| OZ060848.1 | 15 | 21.87 | 43 |
| OZ060849.1 | 16 | 20.45 | 44.5 |
| OZ060850.1 | 17 | 18.2 | 45 |
| OZ060851.1 | 18 | 16.97 | 46 |
| OZ060852.1 | 19 | 16.58 | 45.5 |
| OZ060853.1 | 20 | 13.54 | 47 |
| OZ060854.1 | 21 | 12.42 | 48 |
| OZ060855.1 | 22 | 12.17 | 47 |
| OZ060856.1 | 23 | 9.76 | 47.5 |
| OZ060857.1 | 24 | 7.83 | 49 |
| OZ060858.1 | 25 | 7.8 | 50 |
| OZ060859.1 | 26 | 7.09 | 52 |
| OZ060860.1 | 27 | 6.81 | 51.5 |
| OZ060861.1 | 28 | 6.38 | 52 |
| OZ060862.1 | 29 | 6.01 | 49 |
| OZ060863.1 | 30 | 4.19 | 56 |
| OZ060864.1 | 31 | 3.4 | 57.5 |
| OZ060865.1 | 32 | 2.01 | 46 |
| OZ060866.1 | 33 | 1.98 | 57 |
| OZ060867.1 | 34 | 1.34 | 54.5 |
| OZ060868.1 | 35 | 1.28 | 46 |
| OZ060869.1 | 36 | 1.03 | 60.5 |
| OZ060870.1 | 37 | 0.25 | 55 |
| OZ060871.1 | 38 | 0.22 | 64 |
| OZ060872.1 | 39 | 0.12 | 65 |
| OZ060836.1 | Z | 89.14 | 40.5 |
| OZ060873.1 | MT | 0.02 | 49.5 |

## Genome assembly, curation and evaluation

### *Assembly*

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts ($k$ = 31) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were first assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) using the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

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.

### *Assembly curation*

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. Flat files and maps used in curation were generated via the TreeVal pipeline (Pointon *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation.

### *Assembly quality assessment*

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used to evaluate *k*-mer completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases ($k$ = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed in the blobtoolkit pipeline, a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools (Danecek *et al.*, 2021) and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoaT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these outputs into a blobdir for visualisation.

The blobtoolkit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

## 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

**Table 4. Software tools: versions and sources.**

| Software tool | Version | Source |
|---|---|---|
| BLAST | 2.14.0 | ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/ |
| BlobToolKit | 4.3.9 | https://github.com/blobtoolkit/blobtoolkit |
| BUSCO | 5.5.0 | https://gitlab.com/ezlab/busco |
| bwa-mem2 | 2.2.1 | https://github.com/bwa-mem2/bwa-mem2 |
| fasta_windows | 0.2.4 | https://github.com/tolkit/fasta_windows |
| FastK | 666652151335353eef2fcd58880bcef5bc2928e1 | https://github.com/thegenemyers/FASTK |
| Gfastats | 1.3.6 | https://github.com/vgl-hub/gfastats |
| GoaT CLI | 0.2.5 | https://github.com/genomehubs/goat-cli |
| Hifiasm | 0.19.5-r587 | https://github.com/chhylp123/hifiasm |
| HiGlass | 44086069ee7d4d3f6f3f0012569789ec138f42b8 4aa44357826c0b6753eb28de | https://github.com/higlass/higlass |
| MerquryFK | d00d98157618f4e8d1a9190026b19b471055b 22e | https://github.com/thegenemyers/MERQURY.FK |
| Minimap2 | 2.24-r1122 | https://github.com/lh3/minimap2 |
| MitoHiFi | 3 | https://github.com/marcelauliano/MitoHiFi |
| MultiQC | 1.14, 1.17, and 1.18 | https://github.com/MultiQC/MultiQC |
| Nextflow | 23.10.0 | https://github.com/nextflow-io/nextflow |
| PretextView | 0.2.5 | https://github.com/sanger-tol/PretextView |
| PretextSnapshot | - | https://github.com/sanger-tol/PretextSnapshot |
| purge_dups | 1.2.5 | https://github.com/dfguan/purge_dups |
| samtools | 1.19.2 | https://github.com/samtools/samtools |
| sanger-tol/ascc | 0.1.0 | https://github.com/sanger-tol/ascc |
| sanger-tol/ blobtoolkit | 0.5.0 | https://github.com/sanger-tol/blobtoolkit |
| Seqtk | 1.3 | https://github.com/lh3/seqtk |
| Singularity | 3.9.0 | https://github.com/sylabs/singularity |
| TreeVal | 1.2.0 | https://github.com/sanger-tol/treeval |
| YaHS | 1.2a.2 | https://github.com/c-zhou/yahs |

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 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: Bucephala clangula (common goldeneye). Accession number PRJEB68286; https://identifiers.org/ena.embl/PRJEB68286. The genome sequence is released openly for reuse. The *Bucephala clangula* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745) and the Vertebrate Genomes Project (PRJNA489243). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

## Author information

Members of the Natural History Museum Genome Acquisition Lab are listed here: https://doi.org/10.5281/zenodo.12159242.

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: https://doi.org/10.5281/zenodo.12162482.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: https://doi.org/10.5281/zenodo.12165051.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: https://doi.org/10.5281/zenodo.12160324.

Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.12205391.

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

## Acknowledgements

We would like to thank members of the WWT Living Collections teams and WWT Reserve teams for their help in procuring these samples.

## References

Åhlund M, Andersson M: **Female Ducks can double their reproduction.** *Nature.* 2001; **414**(6864): 600–601.
PubMed Abstract | Publisher 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

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
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: **Common Goldeneye *Bucephala clangula* - BirdLife species factsheet**. 2018.
Reference Source

British Trust for Ornithology: **Goldeneye**. 2015.
Reference Source

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoaT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 24.
PubMed Abstract | Publisher Full Text | Free Full Text

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

da Veiga Leprevost F, Grüning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software**

**standardization.** *Bioinformatics.* 2017; **33**(16): 2580–2582.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023.
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

Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation.** *Genome Biol.* 2023; **24**(1): 74.
PubMed Abstract | Publisher Full Text | Free 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

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text | Free 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

Harry E: **PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps**. 2022.
Reference Source

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Humble K, McGill M: **Watching waterbirds with kate humble and Martin McGill**. Bloomsbury Wildlife, 2019.
Reference Source

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

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey 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

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2, [Accessed 2 April 2024].
Reference Source

Narváez-Gómez JP, Mbye H, Oatley G, *et al.*: **Sanger Tree of Life sample homogenisation: Covaris cryoPREP® Automated Dry Pulverizer.** *protocols.io.* 2023.
Publisher Full Text

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io.* 2023.
Publisher Full Text

Owen M, Atkinson-Willes G, Salmon DG: **Wildfowl in Great Britain: the conservation of wildfowl and their habitats**. 1986.

Pointon DL, Eagles W, Sims Y, *et al.*: **sanger-tol/treeval v1.0.0 – Ancient Atlantis.** 2023.
Publisher Full Text

Pöysä H, Pöysä S: **Nest-Site limitation and density dependence of reproductive output in the common Goldeneye *Bucephala clangula*: implications for the management of cavity-nesting Birds.** *J Appl Ecol.* 2002; **39**(3): 502–510.
Publisher Full Text

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand 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.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

RSPB: **Goldeneye Duck facts *Bucephala clangula***. 2023.
Reference Source

Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.
Publisher Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova 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

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

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text

# Open Peer Review

## Current Peer Review Status: ✔ ✔

**Version 1**

Reviewer Report 30 June 2025

https://doi.org/10.21956/wellcomeopenres.26773.r125092



✔ **Martin Pippel**
Department of Cell and Molecular Biology, Uppsala University, Husargatan 3, Uppsala, Uppsala, Sweden

"The genome sequence of the Common Goldeneye, *Bucephala clangula* (Linnaeus, 1758)" from Colom *et al.* describes a high-quality genome assembly based on PacBio HiFi reads and Hi-C data.

The data note is well structured and the writing is clear and concise. The background information provides a brief insight about the Common Goldeneye. With the described methods and the publicly available sequencing data the assembly should be reproducible by the scientific community.

Methods:

Sample and library preparation are described in great detail, whereas the assembly part falls a bit short. For a genome that possesses microchromosomes it would be nice to make use of the tool MicroFinder (https://github.com/sanger-tol/MicroFinder) in order to identify and assign smaller contigs with mito genes to putative microchromosomes. Furthermore, program arguments of the used tools are not described. Even if all tools were run in default mode, it would be worth mentioning.

I do only have some minor comments and suggestions:

- Table 1: accession ERR12257416 has a read and a base count of 0. ENA only provides a 18Gb bam file for this accession. But, does this data set even belong to the Common Goldeneye and was it used in the assembly?
- Table 2 and text: The final primary assembly consists of 537 scaffolds. But the accession GCA_964059595.1 lists 538 scaffolds. I assume this is due to adding the mitochondrial genome afterwards.
- Figure 3: the interactive version is not available, the phylum is not plotted in the figure – it is just labelled as blank
- Figure 4: the interactive cumulative sequence plot links to the snail plot (Figure 2), the

phylum is not plotted in the figure – it is just labelled as blank
- Assembly:
  - A k-mer database was generated from filtered reads using FastK. How and why were the reads filtered?
  - Yahs was used with the –break option. To my knowledge there is no break option in yahs. By default yahs is doing a contig and a scaffold error correction. The contigs were indeed split into smaller pieces which helps to correct for errors in the contigging. On the other hand yahs can also oversplit contigs. In this Common Goldeneye assembly 25 contigs are smaller or equal to 5 Kb which were trimmed off by yahs from larger contigs. Usually the Hi-C sequencing coverage (i.e. the resolution) is not high enough to support cuts of these short lengths. In some cases contig ends that show telomere motifs are trimmed off – this probably happened for CAXKXP010000212.1 and CAXKXP010000339.1
- Two different spellings for: MERQUERY.FK and Merquery.FK
- Table 4:
  - PretextSnapShot: missing version
  - FastK: I could not find the provided version
  - HiGlass: I could not find the provided versions

**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:*** Genome assembly and annotation, pipeline development

**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.**

Reviewer Report 19 June 2025

https://doi.org/10.21956/wellcomeopenres.26773.r124392

**Rauri C.K. Bowie**
University of California Berkeley, Berkeley, California, USA

This paper utilizes PacBio and Hi-C sequencing to generate a highly contiguous genome assembly for the Common Goldeneye. The study is well designed and will serve as an excellent genomic resource for the research community.

**Background**
When introducing the species it would be helpful to add a map depicting the distribution range of the species.
"*The male's eyes stand out due to its dark green (almost black) head, white cheeks, and black-and-white striped wings."* Point out the white cheek is a circle which distinguishes this species from its close relative Barrow's Goldeneye (*B. islandica*) which has a sickle shaped white patch on the cheek.
4th paragraph under Background: Anatidae comprises both diving and dabbling ducks so it is more accurate to state that they can forage to depths of 7 m. The word "to" is missing.

I understand there may be a reason for the chosen format, but it would have been easier to follow if the Methods section had preceded the Genome Sequence Report. While reviewing, I repeatedly noted the need for additional information, which I later found in the

**Methods**
Under 'Sequencing Data,' it may be clearer to state '45X coverage' rather than just '45 coverage.' It would also be helpful to specify which tissues were used for RNA-seq data collection. In the Methods section, I only noted muscle—was this the only tissue sequenced?
Under 'Assembly Quality Metrics' – What constitutes a good QV (overall quality) value? The general reader would benefit from guidance on how to interpret this metric, or alternatively, you could provide the typical range of values it can span.

Sample acquisition – Please check that the voucher specimen number *NHMUK014561651* is correct as it does to seem typical of the NHM catalogue numbers.
Hi-C Sample preparation and crosslinking - "restriction enzyme master mix" – specify the restriction enzymes used.

Mitochondrial genome assembly: for MitoHiFi was a guide mtDNA genome used? If so, please detail this and provide the NCBI accession number.
Figures 3 and 4 do not add much and can be omitted from the main text and moved to the sup docs.

**Data availability**
*Bucephala clangula* should be in italics.

**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 Biology, Ornithology

**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.**