DATA NOTE

The genome sequence of the turbot, *Scophthalmus maximus* (Linnaeus, 1758) (Pleuronectiformes: Scophthalmidae)

[version 1; peer review: 2 approved]

Rachel Brittain[1], Patrick Adkins[1], Kesella Scott-Somme[1], Vengamanaidu Modepali[1], Marine Biological Association Genome Acquisition Lab, 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

[1]The Marine Biological Association, Plymouth, England, UK

v1 **First published:** 21 Oct 2025, **10**:592
https://doi.org/10.12688/wellcomeopenres.25070.1
Latest published: 21 Oct 2025, **10**:592
https://doi.org/10.12688/wellcomeopenres.25070.1

Abstract

We present a genome assembly from an individual *Scophthalmus maximus* (Turbot; Chordata; Actinopteri; Pleuronectiformes; Scophthalmidae). The genome sequence has a total length of 550.28 megabases. Most of the assembly (99.74%) is scaffolded into 22 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 17.73 kilobases. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Scophthalmus maximus; turbot; genome sequence; chromosomal; Pleuronectiformes

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 21 Oct 2025	✓ view	✓ view

1. **Junrou Huang**, Sun Yat-sen University, Guangzhou, China
2. **Jessica Winn**, Stellenbosch University, Stellenbosch, South Africa

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: Brittain R: Investigation, Resources; **Adkins P**: Investigation, Resources; **Scott-Somme K**: Investigation, Resources; **Modepali V**: Writing – Original Draft Preparation;

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: Brittain R, Adkins P, Scott-Somme K *et al.* **The genome sequence of the turbot, *Scophthalmus maximus* (Linnaeus, 1758) (Pleuronectiformes: Scophthalmidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, **10**:592 https://doi.org/10.12688/wellcomeopenres.25070.1

First published: 21 Oct 2025, **10**:592 https://doi.org/10.12688/wellcomeopenres.25070.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Osteoglossocephalai; Clupeocephala; Euteleosteomorpha; Neoteleostei; Eurypterygia; Ctenosquamata; Acanthomorphata; Euacanthomorphacea; Percomorphaceae; Carangaria; Pleuronectiformes; Pleuronectoidei; Scophthalmidae; *Scophthalmus*; *Scophthalmus maximus* (Linnaeus, 1758) (NCBI:txid52904)

Background

The turbot, *Scophthalmus maximus*, is a flatfish that is part of the Scophthalmidae family and holds global significance. It is naturally distributed in the northeast Atlantic Ocean the Black Sea, the Mediterranean Sea, and the Baltic Sea (Aydin *et al.*, 2022; Bouza *et al.*, 2002; Daniels & Watanabe, 2010), making it a species of interest for the fisheries industry, aquaculture, and environmental conservationists in Europe. The Scophthalmidae family consists of at least 38 identified species of flatfish, with the turbot being classified under the genus Scophthalmus (Fricke, 2020).

The turbot is a predator species, mainly feeding on fish, crustaceans, and bivalves, and it is commonly found in brackish waters, inhabiting sandy, rocky, or mixed bottoms down to 70–80 m (Whitehead *et al.*, 1986). In Mediterranean populations, the spawning season occurs between April and August, while in Atlantic areas, it occurs from May to August. Spawning can be induced throughout the year by adjusting rearing temperatures and day-night rhythms (Imsland *et al.*, 2001). Hormone treatments can also be applied to manage advanced spawning in broodstock and to achieve year-round egg production (Nelson & Neira, 2005).

The turbot is an economically important flatfish species. However, wild turbot populations have been threatened by overfishing and anthropogenic activities (Aydin *et al.*, 2019). To meet the increasing demand, significant efforts have been directed towards turbot aquaculture (Iyengar *et al.*, 2000). Initially selected for aquaculture in the 1970s due to its commercial value in the United Kingdom and France, turbot aquaculture began in Scotland and has since expanded to other European countries and China (FAO, 2021). Additionally, several restoration programs have been implemented to enhance the wild population. Although there is limited research on genetic interactions between farmed and wild turbot, there is a theoretical possibility of genetic introgression and admixture. To ensure the conservation and sustainability of turbot populations, understanding genetic structures and continuous biodiversity monitoring are crucial. The genomes of *Scophthalmus maximus* can be an invaluable tool for understanding the genetic basis of important traits such as growth, sex determination, and disease resistance.

A number of genome sequences for *Scophthalmus maximus* have been produced previously, including GCF_022379125.1 (Xu *et al.*, 2022), the current RefSeq assembly. We present a chromosome-level genome sequence for *Scophthalmus maximus*, produced using the Tree of Life pipeline from a specimen collected in Whitsand Bay, Cornwall, UK (Figure 1). This assembly was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Blaxter *et al.*, 2022).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing and RNA sequencing was an adult *Scophthalmus maximus* (specimen ID MBA-211013-004A, ToLID fScoMax1; Figure 1), collected from Whitsand Bay, Cornwall, UK (latitude 50.3253, longitude –4.2363) on 2021-10-13. The specimen was collected by Kesella Scott-Somme, Patrick Adkins and Rachel Brittain and formally identified by Rachel Brittain based on gross morphology. The fish was first anaesthetised and then overdosed using Aquased (2-phenoxyethanol). Destruction of the brain was used as a secondary method to ensure the animal was deceased before tissue sampling took place as in accordance with Schedule 1 methodology under the home office licence. Samples taken from the animal were preserved on dry ice. For the Darwin Tree of Life sampling and metadata approach, refer to Lawniczak *et al.* (2022).

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating

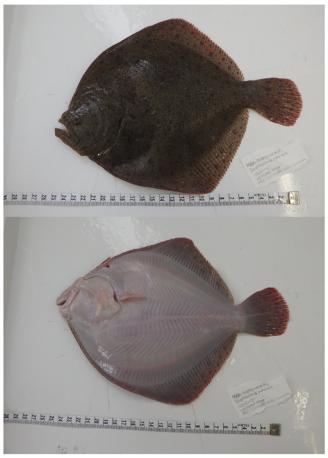

Figure 1. Photographs of the *Scophthalmus maximus* (fScoMax1) specimen used for genome sequencing.

procedures for Darwin Tree of Life barcoding are available on protocols.io.

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The fScoMax1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the gill was homogenised by cryogenic disruption using the Covaris cryoPREP® Automated Dry Pulverizer. HMW DNA was extracted using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). 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 final post-shearing DNA had a Qubit concentration of 19.4 ng/μL and a yield of 2 522.00 ng. The 260/280 spectrophotometric ratio was 1.82, and the 260/230 ratio was 2.74.

RNA was extracted from gill animal tissue of fScoMax1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. 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.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was 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, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen gill tissue of the fScoMax1 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq X to generate 150-bp paired-end reads.

Genome 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 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) with 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. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included 23 breaks, 41 joins, and removal of three haplotypic duplications. This reduced the scaffold count by 15.6%. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) 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 using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all 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), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Scophthalmus maximus* specimen generated 20.53 Gb (gigabases) from 1.92 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 467.38 Mb, with heterozygosity of 0.74% and repeat content of 5.65% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 41× coverage. Hi-C sequencing produced 177.55 Gb from 1 175.80 million reads, which were used

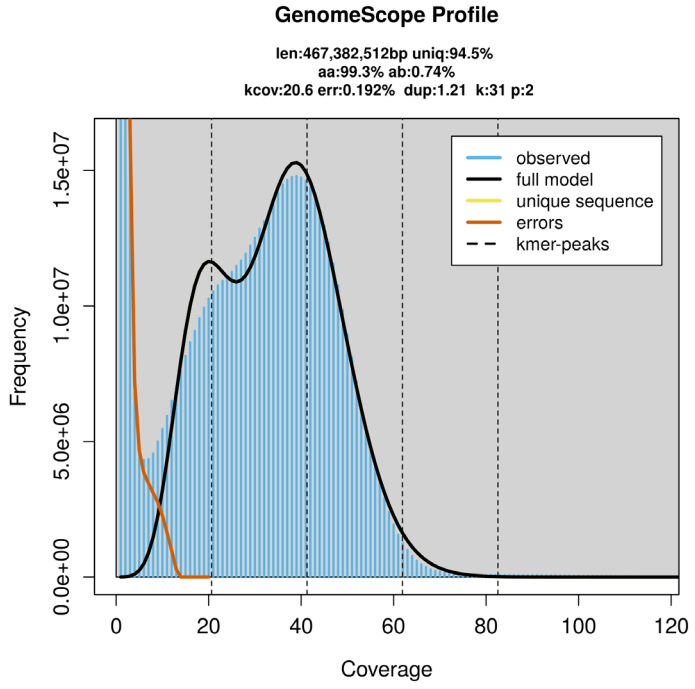

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

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 final assembly has a total length of 550.28 Mb in 107 scaffolds, with 805 gaps, and a scaffold N50 of 26.14 Mb (Table 2).

Most of the assembly sequence (99.74%) was assigned to 22 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

The mitochondrial genome was also assembled (length 17.73 kb, OY978274.1). This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

The combined primary and alternate assemblies achieve an estimated QV of 54.8. The *k*-mer completeness is 93.00% for the primary assembly, 91.54% for the alternate haplotype, and 99.42% for the combined assemblies (Figure 4).

BUSCO v.5.8.3 analysis using the actinopterygii_odb10 reference set (n = 3 640) identified 99.0% of the expected gene set (single = 98.6%, duplicated = 0.4%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in

Table 1. Specimen and sequencing data for BioProject PRJEB69510.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	fScoMax1	fScoMax1	fScoMax1
Specimen ID	MBA-211013-004A	MBA-211013-004A	MBA-211013-004A
BioSample (source individual)	SAMEA110450106	SAMEA110450106	SAMEA110450106
BioSample (tissue)	SAMEA110450979	SAMEA110450985	SAMEA110450979
Tissue	gill animal	gill animal	gill animal
Instrument	Sequel IIe	Illumina NovaSeq 6000	Illumina NovaSeq X
Run accessions	ERR12303950	ERR12318588	ERR12765153
Read count total	1.92 million	1 175.80 million	73.97 million
Base count total	20.53 Gb	177.55 Gb	11.17 Gb

Table 2. Genome assembly statistics.

Assembly name	fScoMax1.1
Assembly accession	GCA_963854745.1
Alternate haplotype accession	GCA_963854755.1
Assembly level	chromosome
Span (Mb)	550.28
Number of chromosomes	22
Number of contigs	912
Contig N50	1.41 Mb
Number of scaffolds	107
Scaffold N50	26.14 Mb
Organelles	Mitochondrion: 17.73 kb

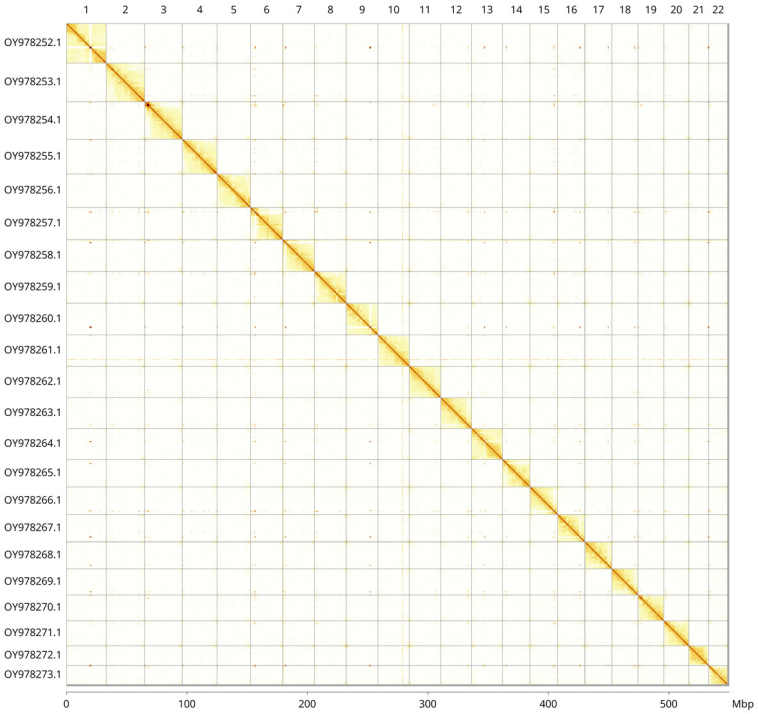

Figure 3. Hi-C contact map of the *Scophthalmus maximus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Scophthalmus maximus* fScoMax1.

INSDC accession	Molecule	Length (Mb)	GC%
OY978252.1	1	33.06	43.50
OY978253.1	2	32.03	43.50
OY978254.1	3	31.24	43
OY978255.1	4	28.77	43.50
OY978256.1	5	27.78	43
OY978257.1	6	26.79	43
OY978258.1	7	26.43	43
OY978259.1	8	26.30	43.50
OY978260.1	9	26.18	43
OY978261.1	10	26.14	43
OY978262.1	11	26.04	43.50
OY978263.1	12	25.68	43.50
OY978264.1	13	25.50	44
OY978265.1	14	22.94	43
OY978266.1	15	22.86	43.50
OY978267.1	16	22.68	43.50
OY978268.1	17	22.41	43.50
OY978269.1	18	21.59	43.50
OY978270.1	19	21.49	44
OY978271.1	20	20.78	43.50
OY978272.1	21	16.36	44
OY978273.1	22	15.80	43.50

Figure 6 shows the distribution of scaffolds by GC proportion and coverage.

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **6.C.Q53**, meeting the recommended reference standard.

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'**,

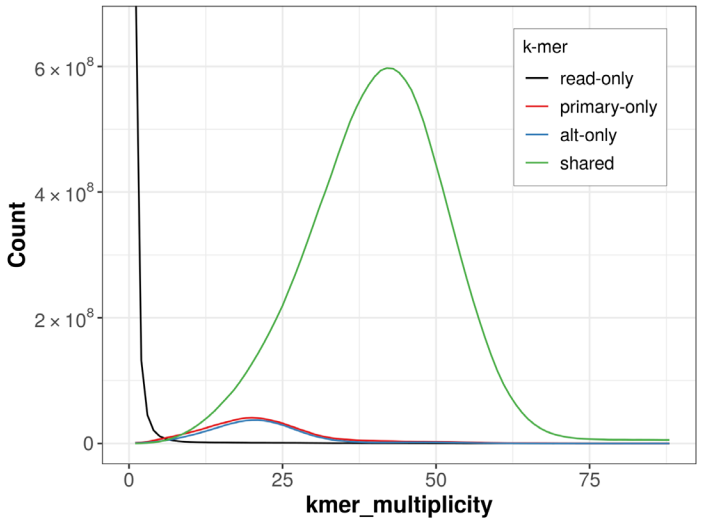

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

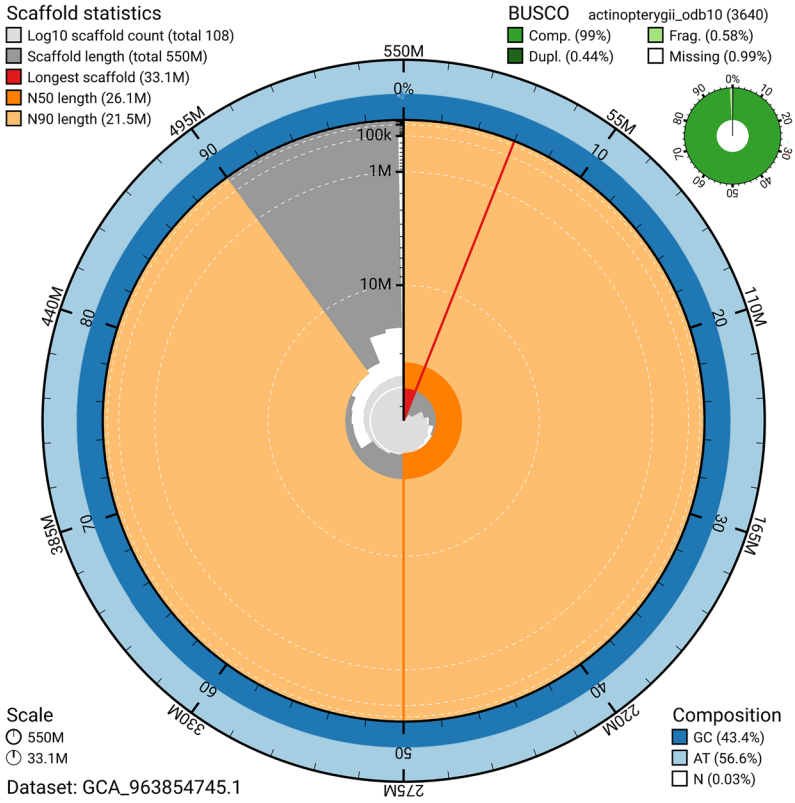

Figure 5. Assembly metrics for fScoMax1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks 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 actinopterygii_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

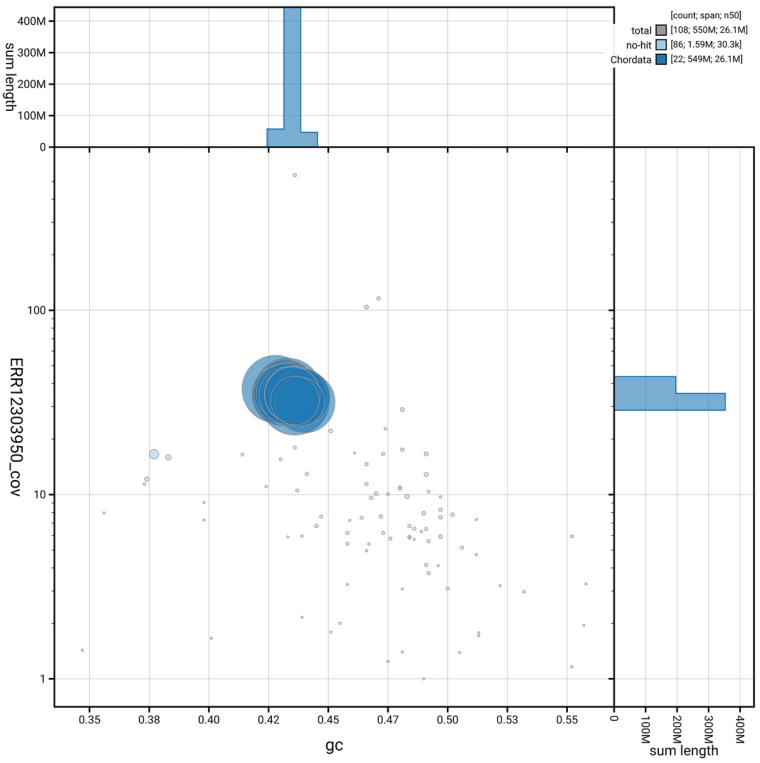

Figure 6. BlobToolKit GC-coverage plot for fScoMax1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth Biogenome Project summary metrics for the *Scophthalmus maximus* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q53	6.C.Q40
Contig N50 length	1.41 Mb	≥ 1 Mb
Scaffold N50 length	26.14 Mb	= chromosome N50
Consensus quality (QV)	Primary: 53.8; alternate: 55.3; combined: 54.8	≥ 40
k-mer completeness	Primary: 93.00%; alternate: 91.54%; combined: 99.42%	≥ 95%
BUSCO	C:99.0% [S:98.6%; D:0.4%]; F:0.6%; M:0.4%; n:3 640	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	99.74%	≥ 90%

which can be found in full on the Darwin Tree of Life website. 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 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: Scophthalmus maximus (turbot). Accession number PRJEB69510. The genome sequence is released openly for reuse. The *Scophthalmus maximus* 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.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the Marine Biological Association Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.4.6	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.8.3	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
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	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi

Software	Version	Source
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.8.0	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

References

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

Aydin I, Polat H, Sahin T: **Reproductive performance of wild and hatchery-reared Black Sea turbot, *Psetta maxima*, in the southern Black Sea coast.** *Turk J Fish Aquat Sci.* 2019; **20**(5): 351–57.
Publisher Full Text

Aydin İ, Öztürk RÇ, Polat H: **Sperm characteristics of wild-caught and hatchery-reared turbot, *Scophthalmus maximus*, originated from the Black Sea.** *J Appl Ichthyol.* 2022; **38**(1): 73–83.
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

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **Sequence locally, think globally: the Darwin Tree of Life Project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
PubMed Abstract | Publisher Full Text | Free Full Text

Bouza C, Presa P, Castro J, *et al.*: **Allozyme and microsatellite diversity in natural and domestic populations of turbot (*Scophthalmus maximus*) in comparison with other Pleuronectiformes.** *Can J Fish Aquat Sci.* 2002; **59**(9): 1460–73.
Publisher Full Text

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, 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

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial Arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
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

Daniels H, Watanabe W: **Practical flatfish culture and stock enhancement**. 2010.
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

FAO: **Fishery and aquaculture statistics.** Global aquaculture production 1950–2019 (FishstatJ), 2021.
Reference Source

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

Fricke R: **Species by family/subfamily in Eschmeyer's catalog of fishes**. 2020.
Reference Source

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–98.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Imsland AK, Foss A, Nævdal G, *et al.*: **Selection or adaptation: differences in growth performance of juvenile turbot (*Scophthalmus maximus* Rafinesque) from two close-by localities off Norway.** *Sarsia.* 2001; **86**(1): 43–51.
Publisher Full Text

Iyengar A, Piyapattanakorn S, Heipel DA, *et al.*: **A suite of highly polymorphic microsatellite markers in turbot (*Scophthalmus maximus* L.) with potential for use across several flatfish species.** *Mol Ecol.* 2000; **9**(3): 368–71.
PubMed Abstract | 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.
Reference Source

Nelson FD, Neira R: **Biotechnology applied to aquaculture i. Classic biotechnologies applied to the reproduction of cultivated species.** *Ciencia e investigación agraria.* 2005; **32**.
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

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI Taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life Project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | Publisher Full Text | Free 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

Whitehead PJP, Bauchot ML, Hureau JC, *et al.*: **Fishes of the North-eastern Atlantic and the Mediterranean: poissons de l'Atlantique Du Nord-Est Et de La Méditerranée.** UNESCO, 1986.
Reference Source

Xu XW, Zheng W, Meng Z, *et al.*: **Identification of stress-related genes by co-expression network analysis based on the improved turbot genome.** *Sci Data.* 2022; **9**(1): 374.
PubMed Abstract | Publisher Full Text | Free 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 06 January 2026

https://doi.org/10.21956/wellcomeopenres.27637.r141840

✔ **Jessica Winn**
Stellenbosch University, Stellenbosch, South Africa

Summary

This manuscript presents a high-quality chromosome-level genome assembly for the turbot (*Scophthalmus maximus*), a commercially important European flatfish species. The assembly was generated using PacBio HiFi long reads (20.53 Gb, ~41× coverage) scaffolded with Hi-C data (177.55 Gb), resulting in 550.28 Mb genome with 22 chromosomal pseudomolecules representing 99.74% of the assembly. The work achieves excellent quality metrics (QV 54.8, 99.0% BUSCO completeness) and meets Earth BioGenome Project reference standards (6.C.Q53). This assembly was produced as part of the Darwin Tree of Life Project.

Strengths

The manuscript demonstrates exemplary best practices in several areas:

- Thorough documentation: Comprehensive software table (Table 5) with versions and sources
- Reproducibility: Protocol.io links, GitLab curation documentation, and public data availability
- Quality metrics: Thorough QC including BUSCO, k-mer completeness, QV scores, and EBP benchmarking
- Transparency: Detailed curation steps quantified (23 breaks, 41 joins, 3 haplotypic duplications removed)
- High-quality assembly: Excellent contiguity (N50 26.14 Mb) and completeness
- Clear presentation: Well-organized tables and informative figures
- Ethical compliance: Thorough legal/governance section and sampling protocols

Key issues

1. Scophthalmidae species count

Background (paragraph 1): "The Scophthalmidae family consists of at least 38 identified species of flatfish"

This statement is misleading. While Fricke (2020) does list 38 available species names for Scophthalmidae, only 9 are taxonomically valid species distributed across 4 genera. The remaining names likely represent synonyms, subspecies, or historically recognized taxa no longer considered

valid. The current wording incorrectly suggests 38 distinct valid species exist. Reword to accurately reflect taxonomic validity and provide some more phylogenetic context to prevent confusion about the family's diversity.
2. Incomplete results reporting
Genome sequence report: Methods mentions ASCC decontamination pipeline was run, but Results never reports the outcome.
Add one sentence stating either that no contamination was detected during ASCC screening or state how many contaminant scaffolds and bases were identified and removed.

Minor suggestions

Background section
1. Missing punctuation
Paragraph 1: "in the northeast Atlantic Ocean the Black Sea..."
Add comma: "in the northeast Atlantic Ocean*,* the Black Sea..."
2. Vague language
Paragraph 1: "holds global significance" is imprecise.
Suggestion: "is of significant economic and ecological importance" or "is commercially and ecologically important"
3. Awkward phrasing
Paragraph 3: "Initially selected for aquaculture in the 1970s due to its commercial value in the United Kingdom and France, turbot aquaculture began in Scotland and has since expanded..."
Suggestion: "Turbot was initially selected for aquaculture in Scotland in the 1970s due to its high commercial value in European markets, and production has since expanded to other European countries and China."

Methods section
4. Table 5 reference placement
Table 5 (comprehensive software versions) is mentioned only in the Data availability statement but should be referenced in the Methods.
Suggestion: Add to first paragraph of Methods: "All software tools and versions used in this study are listed in Table 5". Otherwise move this table into the Methods section.
5. RNA-seq fragment size clarity
"Libraries were size-selected to produce fragments between 100-300 bp" seems quite small.
Clarify whether this refers to cDNA library insert size or final library size including adapters.
6. Hi-C pooling specificity
"Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools"
This sentence is unclear. Specify which approach was used (equimolar OR weighted, not both). If different pools used different strategies, this should be specified.

Genome assembly report section:
7. RNA-seq purpose
Sequence data: "RNA sequencing data were also generated and are available in public sequence repositories."
Suggestion: Consider adding a forward reference: "RNA sequencing data were generated for genome annotation and are available in public repositories (see Data Availability)."
8. Hi-C quality metrics
Sequence data: Hi-C data volume is reported but not quality metrics.
Consider adding the percentage of valid pairs and long-range contacts with the read number to validate scaffolding quality if available.
9. Gap characterization
Assembly statistics: "805 gaps" is mentioned but not explained.

Add one sentence clarifying these are Hi-C scaffold gaps (inter-contig junctions) and provide average estimated size if available, e.g.: 'The 805 gaps represent junctions between contigs within Hi-C scaffolds, with an average estimated size of ~X bp.'"
10. Curation impact clarity
Methods quantify curation (23 breaks, 41 joins, 3 haplotypic duplications) and states "15.6% reduction" but the genome assembly report doesn't connect these.
Suggestion: Expand on this in the assembly report. "Manual curation reduced the initial scaffold count from X to 107 (15.6% reduction) through 23 breaks, 41 joins, and removal of three haplotypic duplications."
Conclusion
This is a high-quality genome paper that demonstrates exemplary documentation and reproducibility practices. The assembly quality is excellent and meets international standards. With correction of some minor issues, the manuscript will be scientifically sound and ready to be indexed.

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

Are the protocols appropriate and is the work technically sound?
Partly

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

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research expertise is in phylogenomics, population genomics, and molecular ecology, with experience in bioinformatics pipelines and genome analysis. While I have assembled mitogenomes and worked with genomic data, I have not personally conducted whole-genome assembly using PacBio HiFi reads. I evaluated this manuscript based on my genomic data analysis experience, comparison with published genome assemblies, and review of relevant literature, though I may lack some technical nuances specific to long-read assembly workflows.

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 24 November 2025

https://doi.org/10.21956/wellcomeopenres.27637.r139494

Junrou Huang
Sun Yat-sen University, Guangzhou, China

This manuscript presents a high-quality, chromosome-level genome assembly for the turbot, generated under the Darwin Tree of Life project. The assembly demonstrates excellent quality, with 99.74% of the sequence anchored into 22 chromosomes, high BUSCO completeness, and robust quality scores. The integration of PacBio HiFi, Hi-C, and RNA-seq data, followed by manual curation, has resulted in a reference-grade genome that meets the standards of the Earth BioGenome Project. The data are comprehensively documented and made readily available to the community. This resource will be invaluable for research in areas such as aquaculture, comparative genomics, and conservation biology.

I recommend indexing of this manuscript.

The study is technically sound and the data represent a significant contribution. However, the Introduction would be strengthened by explicitly stating the rationale for generating a new genome assembly, given the prior existence of a chromosome-level assembly (GCF_022379125.1). A concise justification would better contextualize the necessity and added value of this work.

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

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, Comparative genomics, Single-cell transcriptomics

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.