





## DATA NOTE

# The genome sequence of the common sole, *Solea solea*

(Linnaeus, 1758)

[version 1; peer review: 2 approved]

Enora Geslain <sup>1</sup>, Filip A.M. Volckaert <sup>1</sup>, Ann M. Mc Cartney<sup>2</sup>,

Giulio Formenti <sup>3,4</sup>, Alice Mouton <sup>5,6</sup>,

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

<sup>1</sup>Laboratory of Biodiversity and Evolutionary Genomics, KU Leuven, Leuven, Belgium

<sup>2</sup>Genomics Institute, University of California Santa Cruz, Santa Cruz, California, USA

<sup>3</sup>The Vertebrate Genome Laboratory, The Rockefeller University, New York, New York, USA

<sup>4</sup>Department of Biology, University of Florence, Sesto Fiorentino, Italy

<sup>5</sup>InBios-Conservation Genetics Laboratory, University of Liege, Liege, Belgium

<sup>6</sup>Leibniz Institut für Zoo und Wildtierforschung, Berlin, Germany

**V1** First published: 20 Jan 2025, 10:25  
<https://doi.org/10.12688/wellcomeopenres.23353.1>  
Latest published: 20 Jan 2025, 10:25  
<https://doi.org/10.12688/wellcomeopenres.23353.1>

## Abstract

We present a genome assembly from an individual female *Solea solea* (Linnaeus, 1758) (the common sole; Chordata; Actinopteri; Pleuronectiformes; Soleidae). The genome sequence spans 643.80 megabases. Most of the assembly (97.81%) is scaffolded into 21 chromosomal pseudomolecules. The mitochondrial genome has also been assembled and is 17.03 kilobases in length. Gene annotation of this assembly on Ensembl identified 21,646 protein-coding genes.

## Keywords



*Solea solea*, common sole, genome sequence, chromosomal, Pleuronectiformes




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

## Open Peer Review

### Approval Status

	1	2
<b>version 1</b>		
20 Jan 2025	<a href="#">view</a>	<a href="#">view</a>

1. **Huria Marnis** , National Research and Innovation Agency (BRIN), Cibinong, Indonesia

2. **Shigehiro Kuraku**, National Institute of Genetics, Shizuoka, Japan

Any reports and responses or comments on the article can be found at the end of the article.

**Corresponding author:** Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team ([Mark.Blaxter@sanger.ac.uk](mailto:Mark.Blaxter@sanger.ac.uk))

**Author roles:** **Geslain E:** Data Curation, Formal Analysis, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Volckaert FAM:** Conceptualization, Investigation, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; **Mc Cartney AM:** Conceptualization, Project Administration, Supervision; **Formenti G:** Conceptualization, Project Administration, Supervision; **Mouton A:** Conceptualization, Project Administration, Supervision;

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by Wellcome through core funding to the Wellcome Sanger Institute [206194, <https://doi.org/10.35802/206194>].

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

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

**How to cite this article:** Geslain E, Volckaert FAM, Mc Cartney AM *et al.* **The genome sequence of the common sole, *Solea solea* (Linnaeus, 1758) [version 1; peer review: 2 approved]** Wellcome Open Research 2025, 10:25 <https://doi.org/10.12688/wellcomeopenres.23353.1>

**First published:** 20 Jan 2025, 10:25 <https://doi.org/10.12688/wellcomeopenres.23353.1>

## Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Osteoglossocephalai; Clupeocephala; Euteleostomorpha; Neoteleostei; Eurypterygia; Ctenosquamata; Acanthomorphata; Euacanthomorphacea; Percomorphaceae; Carangaria; Pleuronectiformes; Pleuronectoidei; Soleidae; *Solea*; *Solea solea*, (Linnaeus, 1758) (NCBI:txid90069).

## Background

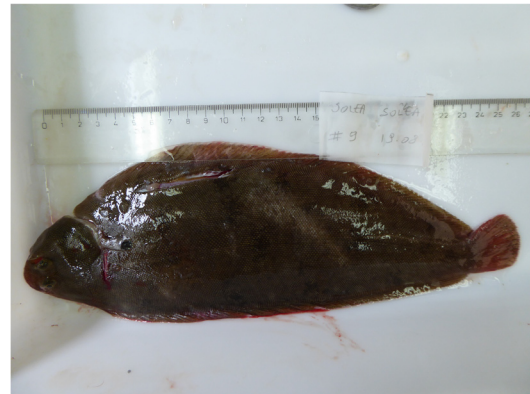
The demersal flatfish *Solea solea* (Linnaeus, 1758), known commonly as common sole, typically burrows in sandy and muddy bottoms at depths of less than 150 m. This species is widespread in warm and cold temperate seas, including East Atlantic continental shelf waters from Trondheim Fjord (65° N; Norway) southward, the Mediterranean, including the Sea of Marmara, Bosphorus and southwestern Black Sea, and the northwestern African coastal waters southward to Senegal, including Cape Verde (5° N) (Froese & Pauly, 2024). Common sole reach sizes of max. 70 cm, but more commonly 15 to 45 cm and may reach 40 years of age (ICES, 2006). Planktonic larvae feed on copepod nauplii; juveniles and adults feed nocturnally on benthic invertebrates such as polychaetes, siphons of bivalves and small crustaceans (amphipods) and young stages of echinoderms (ICES, 2006). Common sole spawn along the coast in proximity of estuaries; postlarvae settle on inshore nursery grounds where they grow to subadults during two to three years. Recruitment is highly variable.

Common sole is a high value consumption flatfish with a well-known biology and limited genomic resources (Gibson *et al.*, 2014). It is targeted by commercial fisheries using beam trawling, formerly electrotrawling, and to a lesser extent gill netting (ICES, 2006; Rijnsdorp *et al.*, 2024). Major fished stocks are managed regionally by ICES without evidence for a mismatch with genetic structure (Diopere *et al.*, 2018). Total landings of 87,120 metric tonnes were estimated in 1992 down to 31,030 metric tonnes in 2019 (Pauly *et al.*, 2020). The International Union for the Conservation of Nature has listed common sole as “Data Deficient” because of the poor availability of data over sections of the species range (Tous *et al.*, 2015).

Here we present, to our knowledge, the first complete chromosomal-level genome sequence reported for *Solea solea*, based on a female specimen from the Kwintebank (51.283 N; 2.65 E), North Sea, Belgium, kept at the Belgian Institute of Natural Sciences, Brussels, Belgium (voucher numbers KBIN/IRSNB/RBINS 27235 and KBIN/IRSNB/RBINS AB42614133).

## Genome sequence report

The genome of an adult female *Solea solea* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating a total of 20.96 Gb (gigabases) from 2.27 million reads, providing approximately 32-fold coverage.



**Figure 1. Photograph of the adult female *Solea solea* (fSolSol10) specimen used for genome sequencing.**

Primary assembly contigs were scaffolded with chromosome conformation Hi-C data, which produced 130.51 Gb from 864.29 million reads, yielding an approximate coverage of 203-fold. Specimen and sequencing information is summarised in Table 1.

Assembly errors were corrected during manual curation: including 22 missing joins or mis-joins and three haplotypic duplications. This reduced the scaffold number by 2.8% and increased the scaffold N50 by 0.21%. The final assembly has a total length of 643.80 Mb in 242 sequence scaffolds with a scaffold N50 of 29.1 Mb (Table 2). The total count of gaps in the scaffolds is 373.

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 (97.81%) was assigned to 21 chromosomal-level scaffolds. Chromosome-scale scaffolds confirmed by the Hi-C data are named in order of size (Figure 5; Table 3). While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited. The mitochondrial genome was also assembled and can be found as a contig within the multifasta file of the genome submission.

The estimated Quality Value (QV) of the final assembly is 60.9 with *k*-mer completeness of 99.36% (combined assemblies), and the assembly has a BUSCO v5.3.2 completeness of 98.3% (single = 97.2%, duplicated = 1.1%), using the actinopterygii\_odb10 reference set (*n* = 3,640). The assembly achieves the Earth BioGenome Project reference standard of 6.C.Q61, thus exceeding the minimum reference standard of 6.C.Q40. Other quality metrics are given in Table 2.

**Table 1. Specimen and sequencing data for *Solea solea*.**

Project information			
Study title	Solea solea (common sole)		
Umbrella BioProject	PRJEB61337		
Species	<i>Solea solea</i>		
BioSample	SAMEA10984647		
NCBI taxonomy ID	90069		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	fSolSol10	SAMEA10984681	gonad
Hi-C sequencing	fSolSol10	SAMEA10984680	gill animal
RNA sequencing	fSolSol7	SAMEA10984669	gonad
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR11242545	8.64e+08	130.51
PacBio Sequel IIe	ERR11242127	2.27e+06	20.96
RNA Illumina NovaSeq 6000	ERR12245556	7.92e+07	11.96
RNA Illumina NovaSeq 6000	ERR11242544	7.51e+07	11.34

## Genome annotation report

The *Solea solea* genome assembly (GCA\_958295425.1) was annotated at the European Bioinformatics Institute (EBI) on Ensembl Rapid Release. The resulting annotation includes 51,064 transcribed mRNAs from 21,646 protein-coding and 1,553 non-coding genes (Table 2; [https://rapid.ensembl.org/Solea\\_solea\\_GCA\\_958295425.1/Info/Index](https://rapid.ensembl.org/Solea_solea_GCA_958295425.1/Info/Index)). The average transcript length is 21,373.58. There are 2.20 coding transcripts per gene and 13.64 exons per transcript.

## Methods

### Sample acquisition

An adult *Solea solea* (specimen ID ERGA\_FV\_BE\_019, ToLID fSolSol10) was collected from Kwinkebank (North Sea; latitude 51.28, longitude 2.65) on 2021-08-19. The specimen was collected, identified and preserved by Filip A.M. Volckaert (KU Leuven). This specimen was used for genome sequencing and Hi-C data for scaffolding.

The specimen used for RNA sequencing (ToLID fSolSol7) was a juvenile specimen collected from Wenduinebank (W03) (North Sea; latitude 51.28, longitude 2.95) on 2021-07-21. The specimen was collected, identified and preserved by Filip A.M. Volckaert (KU Leuven).

### 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 core procedures: sample preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023). The fSolSol10 sample was weighed and dissected on dry ice (Jay *et al.*, 2023) and gonad tissue was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023).

HMW DNA was extracted using the Automated MagAttract v1 protocol (Sheerin *et al.*, 2023). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Todorovic *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.

**Table 2. Genome assembly data for *Solea solea*, fSolSol10.1.**

Genome assembly		
Assembly name	fSolSol10.1	
Assembly accession	GCA_958295425.1	
Accession of alternate haplotype	GCA_958295035.1	
Span (Mb)	643.80	
Number of contigs	616	
Number of scaffolds	242	
Longest scaffold (Mb)	47.06	
Assembly metrics*		Benchmark
Contig N50 length (Mb)	2.6	$\geq 1 \text{ Mb}$
Scaffold N50 length (Mb)	29.1	= chromosome N50
Consensus quality (QV)	60.9	$\geq 40$
k-mer completeness	99.36% (combined assemblies)	$\geq 95\%$
BUSCO v 5.3.2 lineage: actinopterygii_odb10**	C:98.3%,S:97.2%,D:1.1%, F:0.5%,M:1.2%,n:3,640	$S > 90\%$ $D < 5\%$
Percentage of assembly mapped to chromosomes	97.81%	$\geq 90\%$
Sex chromosomes	Not identified	localised homologous pairs
Organelles	Mitochondrial genome: 17.03 kb	complete single alleles
Genome annotation of assembly GCA_958295425.1 at Ensembl		
Number of protein-coding genes	21,646	
Number of non-coding genes	1,553	
Number of gene transcripts	51,064	

\* Assembly metric benchmarks are adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024.

\*\* A full set of BUSCO scores is available at [https://blobtoolkit.genomehubs.org/view/fSolSol10\\_1/dataset/fSolSol10\\_1/busco](https://blobtoolkit.genomehubs.org/view/fSolSol10_1/dataset/fSolSol10_1/busco).

RNA was extracted from gonad tissue of fSolSol7 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 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.

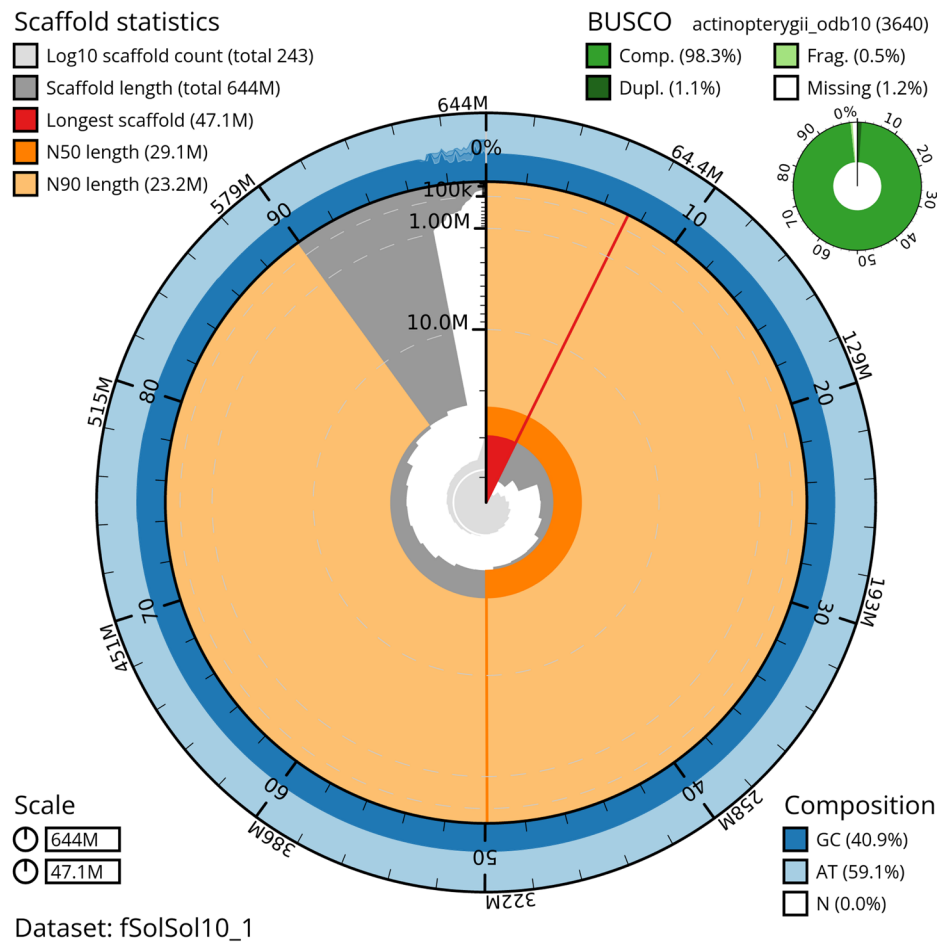
#### Hi-C preparation

Tissue from the gill of the fSolSol10 sample was processed at the WSI Scientific Operations core, using the Arima-HiC v2 kit. Tissue (stored at  $-80^{\circ}\text{C}$ ) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde.

After crosslinking, the tissue was homogenised using the Diagenode Power Masher-II and BioMasher-II tubes and pestles. Following the kit manufacturer's instructions, crosslinked DNA was digested using a restriction enzyme master mix. The 5'-overhangs were then filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation.

#### Library preparation and sequencing

Pacific Biosciences SMRTbell libraries were constructed using the Revio HiFi prep kit, according to the manufacturers'



**Figure 2. Genome assembly of *Solea solea*, fSolSol10.1: metrics.** The BlobToolKit snail plot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 bins around the circumference with each bin representing 0.1% of the 643,774,986 bp assembly. The distribution of scaffold lengths is shown in dark grey with the plot radius scaled to the longest scaffold present in the assembly (47,064,963 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 scaffold lengths (29,068,826 and 23,155,029 bp), respectively. The pale grey spiral shows the cumulative scaffold count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the actinopterygii\_odb10 set is shown in the top right. An interactive version of this figure is available at [https://blobtoolkit.genomehubs.org/view/fSolSol10\\_1/dataset/fSolSol10\\_1/snail](https://blobtoolkit.genomehubs.org/view/fSolSol10_1/dataset/fSolSol10_1/snail).

instructions. DNA sequencing was performed by the Scientific Operations core at the WSI on a Pacific Biosciences Revio instrument.

For Hi-C library preparation, DNA was fragmented to a size of 400 to 600 bp using a Covaris E220 sonicator. The DNA was then enriched, barcoded, and amplified using the NEB-Next Ultra II DNA Library Prep Kit following manufacturers' instructions. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit, following the manufacturer's

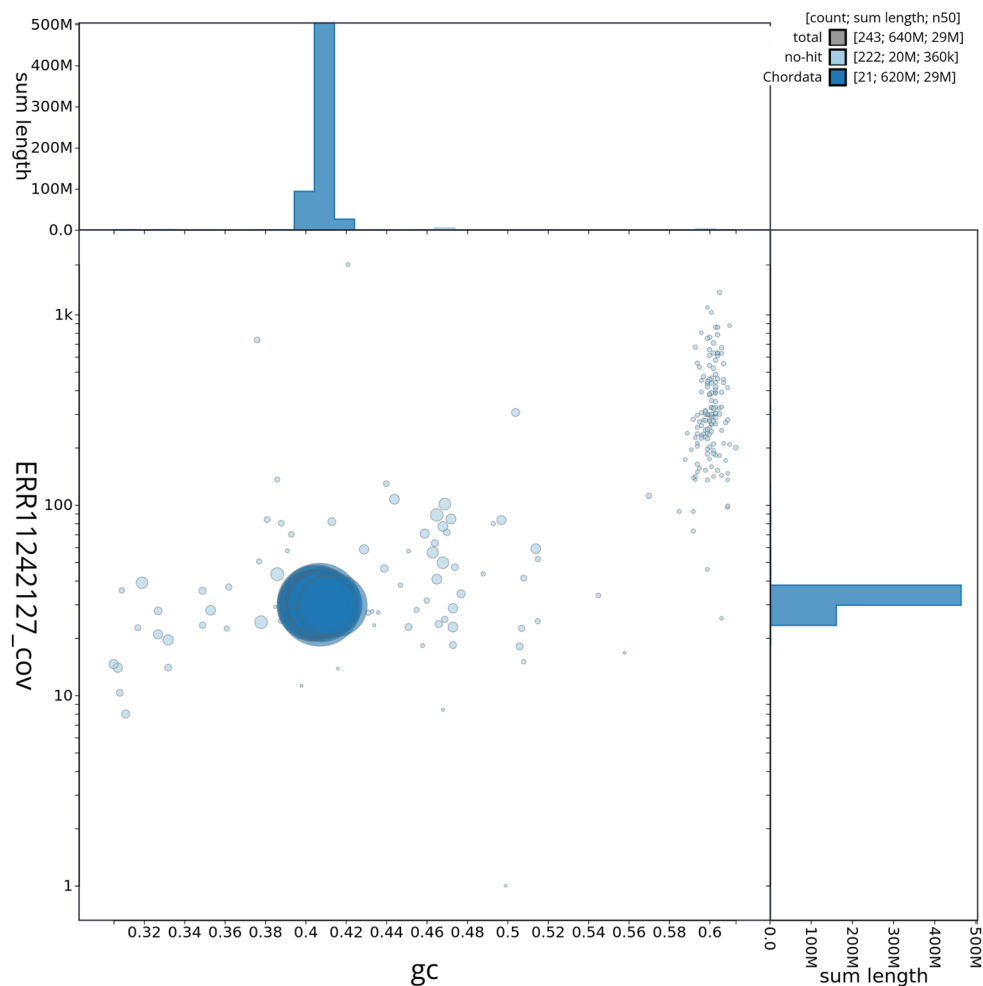
instructions. RNA sequencing was performed on the Illumina NovaSeq 6000 instrument.

### Genome assembly, curation and evaluation

#### Assembly

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 were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) in YaHS (Zhou *et al.*, 2023) using the --break option. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).





**Figure 3. Genome assembly of *Solea solea*, fSolSol10.1: Blob plot of base coverage in ERR11242127 against GC proportion for sequences in assembly fSolSol10.1.** Sequences are coloured by phylum. Circles are sized in proportion to sequence length. Histograms show the distribution of sequence length sum along each axis. An interactive version of this figure is available at [https://blobtoolkit.genomehubs.org/view/fSolSol10\\_1/dataset/fSolSol10\\_1/blob](https://blobtoolkit.genomehubs.org/view/fSolSol10_1/dataset/fSolSol10_1/blob).

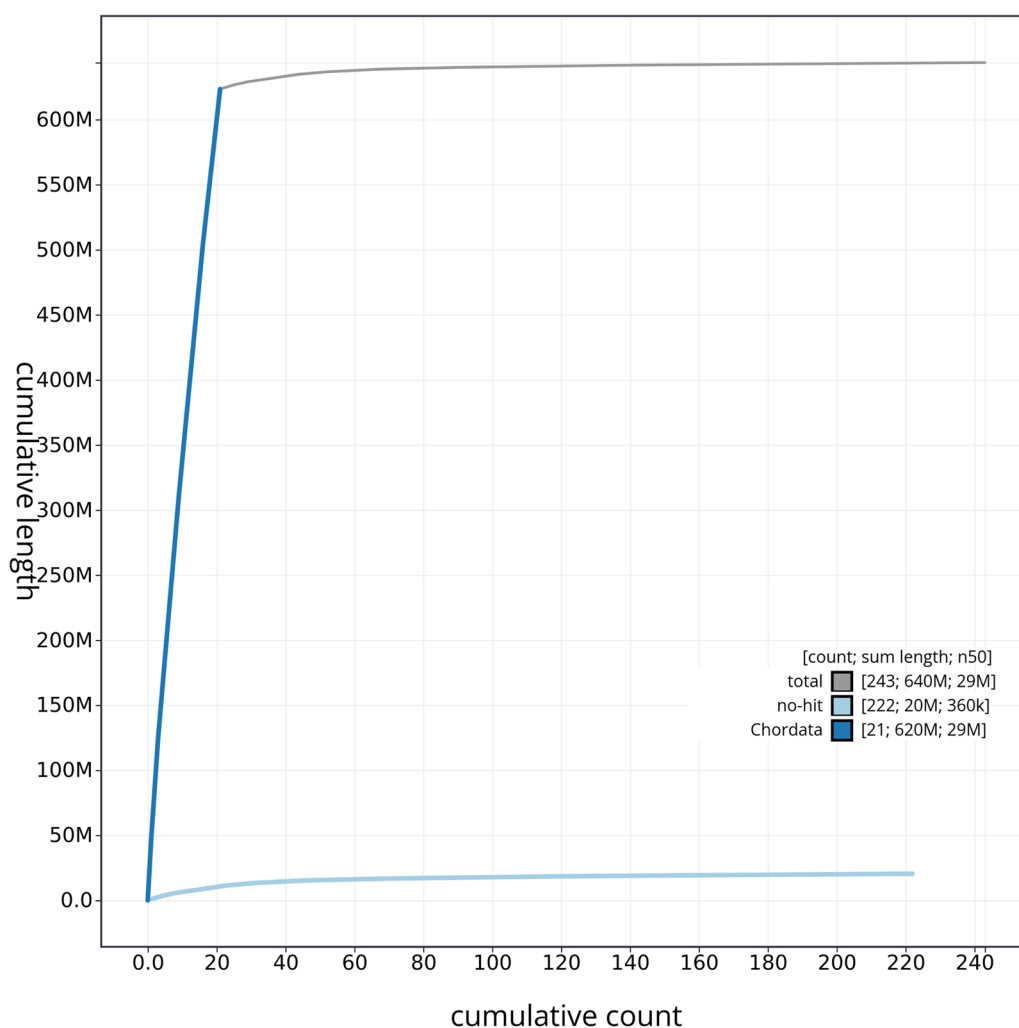
The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

#### Assembly curation

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

#### Evaluation of the final assembly

A Hi-C map for the final assembly was produced using bwa-mem2 (Vasimuddin *et al.*, 2019) in the Cooler file format (Abdennur & Mirny, 2020). To assess the assembly metrics, the *k*-mer completeness and QV consensus quality values were calculated in Merquy (Rhie *et al.*, 2020). This work was done using the “sanger-tol/readmapping” (Surana *et al.*, 2023a) and “sanger-tol/genomenote” (Surana *et al.*, 2023b) pipelines. The genome readmapping pipelines were developed using the nf-core tooling (Ewels *et al.*, 2020), use MultiQC (Ewels *et al.*, 2016), and make extensive use of the Conda package manager, the Bioconda initiative (Grünig *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), and the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions. The genome was also analysed within the BlobToolKit environment (Challis *et al.*, 2020) and BUSCO scores (Manni *et al.*, 2021) were calculated.



**Figure 4. Genome assembly of *Solea solea* fSolSol10.1: BlobToolKit cumulative sequence plot.** The grey line shows cumulative length for all sequences. Coloured lines show cumulative lengths of sequences assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at [https://blobtoolkit.genomehubs.org/view/fSolSol10\\_1/dataset/fSolSol10\\_1/cumulative](https://blobtoolkit.genomehubs.org/view/fSolSol10_1/dataset/fSolSol10_1/cumulative).

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

#### Genome annotation

The Ensembl Genebuild annotation system (Aken *et al.*, 2016) was used to generate annotation for the *Solea solea* assembly (GCA\_958295425.1) in Ensembl Rapid Release at the EBI. Annotation was created primarily through alignment of transcriptomic data to the genome, with gap filling via protein-to-genome alignments of a select set of proteins from UniProt (UniProt Consortium, 2019).

#### Wellcome Sanger Institute – Legal and Governance

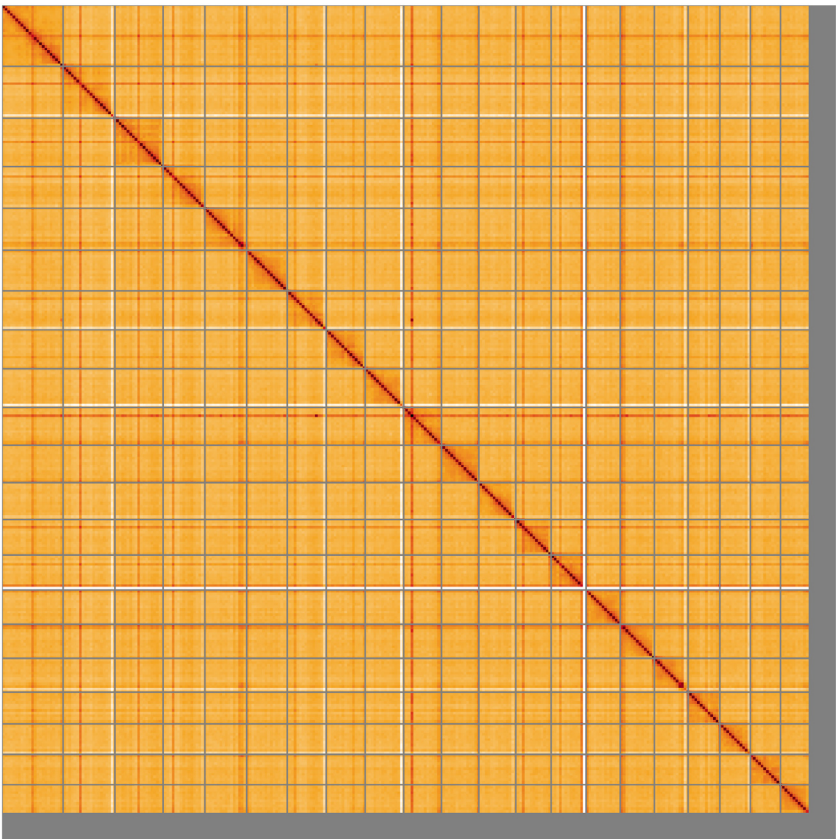
The materials that have contributed to this genome note have been supplied by a Tree of Life collaborator.

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)





**Figure 5. Genome assembly of *Solea solea* fSolSol10.1: Hi-C contact map of the fSolSol10.1 assembly, visualised using HiGlass.** Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at [https://genome-note-higlass.tol.sanger.ac.uk/l/?d=FSUllwVQRfiQQBsuU\\_Y3AQ](https://genome-note-higlass.tol.sanger.ac.uk/l/?d=FSUllwVQRfiQQBsuU_Y3AQ).

**Table 3. Chromosomal pseudomolecules in the genome assembly of *Solea solea*, fSolSol10.**

INSDC accession	Name	Length (Mb)	GC%
OY282534.1	1	47.06	40.5
OY282535.1	2	40.01	40.5
OY282536.1	3	37.3	41.0
OY282537.1	4	26.89	41.0
OY282538.1	5	32.35	40.5
OY282539.1	6	32.19	40.5
OY282540.1	7	31.38	40.5
OY282541.1	8	30.09	40.5
OY282542.1	9	29.98	40.5
OY282543.1	10	29.76	40.5
OY282544.1	11	29.07	40.5
OY282545.1	12	29.01	40.5

INSDC accession	Name	Length (Mb)	GC%
OY282546.1	13	28.33	40.5
OY282547.1	14	27.38	40.5
OY282548.1	15	26.54	40.5
OY282549.1	16	26.39	41.5
OY282550.1	17	25.92	41.0
OY282551.1	18	24.5	41.5
OY282552.1	19	23.71	40.5
OY282553.1	20	23.16	41.0
OY282554.1	21	22.35	41.0
OY282555.1	MT	0.02	44.0

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

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

Software tool	Version	Source
BlobToolKit	4.2.1	<a href="https://github.com/blobtoolkit/blobtoolkit">https://github.com/blobtoolkit/blobtoolkit</a>
BUSCO	5.3.2	<a href="https://gitlab.com/ezlab/busco">https://gitlab.com/ezlab/busco</a>
bwa-mem2	2.2.1	<a href="https://github.com/bwa-mem2/bwa-mem2">https://github.com/bwa-mem2/bwa-mem2</a>
Cooler	0.8.11	<a href="https://github.com/open2c/cooler">https://github.com/open2c/cooler</a>
Gfastats	1.3.6	<a href="https://github.com/vgl-hub/gfastats">https://github.com/vgl-hub/gfastats</a>
Hifiasm	0.16.1-r375	<a href="https://github.com/chhylp123/hifiasm">https://github.com/chhylp123/hifiasm</a>
HiGlass	1.11.6	<a href="https://github.com/higlass/higlass">https://github.com/higlass/higlass</a>
Mercury.FK	d00d98157618f4e8d1a9190026b19b471055b22e	<a href="https://github.com/thegenemyers/MERQURY.FK">https://github.com/thegenemyers/MERQURY.FK</a>
MitoHiFi	3	<a href="https://github.com/marcelauliano/MitoHiFi">https://github.com/marcelauliano/MitoHiFi</a>
PretextView	0.2	<a href="https://github.com/wtsi-hpag/PretextView">https://github.com/wtsi-hpag/PretextView</a>
purge_dups	1.2.5	<a href="https://github.com/dfguan/purge_dups">https://github.com/dfguan/purge_dups</a>
sanger-tol/genomenote	v1.0	<a href="https://github.com/sanger-tol/genomenote">https://github.com/sanger-tol/genomenote</a>
sanger-tol/readmapping	1.1.0	<a href="https://github.com/sanger-tol/readmapping/tree/1.1.0">https://github.com/sanger-tol/readmapping/tree/1.1.0</a>
Singularity	3.9.0	<a href="https://github.com/sylabs/singularity">https://github.com/sylabs/singularity</a>
YaHS	1.2a.2	<a href="https://github.com/c-zhou/yahs">https://github.com/c-zhou/yahs</a>

### Data availability

European Nucleotide Archive: *Solea solea* (common sole). Accession number PRJEB61337; <https://identifiers.org/ena.embl/PRJEB61337>. The genome sequence is released openly for reuse. The *Solea solea* genome sequencing initiative is part of the European Reference Genome Atlas (ERGA) pilot project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Metadata for specimens, BOLD barcode results, spectra estimates, sequencing runs, contaminants and pre-curation assembly statistics are given at <https://links.tol.sanger.ac.uk/species/90069>.

### Author information

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>.

### References

- Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Aken BL, Ayling S, Barrell D, et al.: **The ensembl gene annotation system.** *Database (Oxford).* 2016; **2016**: baw093.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Allio R, Schomaker-Bastos A, Romiguier J, et al.: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target**

- enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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, Grünig 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](#)
- Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023.  
[Publisher Full Text](#)
- Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of syntenic and structural variation.** *Genome Biol.* 2023; **24**(1): 74.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Diopere E, Vandamme SG, Hablützel PI, *et al.*: **Seascape genetics of a flatfish reveals local selection under high levels of gene flow.** *ICES J Mar Sci.* 2018; **75**(2): 675–689.  
[Publisher Full Text](#)
- do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana.** *protocols.io.* 2023.  
[Publisher Full Text](#)
- 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](#)
- Froese R, Pauly D: **FishBase.** 2024.  
[Reference Source](#)
- Gibson RN, Nash RDM, Geffen AJ (eds.), *et al.*: **Flatfishes.** Wiley. 2014.  
[Publisher Full Text](#)
- Grünig 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](#)
- Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): gaa153.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- ICES: **Sole Solea solea.** 2006.  
[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](#)
- Manni M, Berkeley MR, Seppie 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 V.1.** *protocols.io.* 2023.  
[Publisher Full Text](#)
- Pauly D, Zeller D, Palomares MLD (eds.): **Sea around us: concepts, design and data.** 2020.  
[Reference Source](#)
- 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.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rijnsdorp AD, Boute PG, Tian JC, *et al.*: **Electrotrawling can improve the sustainability of the bottom trawl fishery for sole: a review of the evidence.** *Rev Fish Biol Fisher.* 2024; **34**(3): 959–993.  
[Publisher Full Text](#)
- Sheerin E, Sampaio F, Oatley G, *et al.*: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.1.** *protocols.io.* 2023.  
[Publisher Full Text](#)
- Strickland M, Cornwell C, Howard C: **Sanger Tree of Life fragmented DNA clean up: manual SPRI.** *protocols.io.* 2023.  
[Publisher Full Text](#)
- Surana P, Muffato M, Qi G: **Sanger-tol/readmapping: sanger-tol/readmapping v1.1.0 - Hebridean Black (1.1.0).** *Zenodo.* 2023a.  
[Publisher Full Text](#)
- Surana P, Muffato M, Sadasivan Baby C: **sanger-tol/genomenote (v1.0.dev).** *Zenodo.* 2023b.  
[Publisher Full Text](#)
- Todorovic M, Sampaio F, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for PacBio HiFi.** *protocols.io.* 2023.  
[Publisher Full Text](#)
- Tous P, Sidibe A, Mbye E, *et al.*: **Solea solea.** *The IUCN Red List of threatened species.* 2015; **2015**: e.T198739A15595369, [Accessed 17 October 2024].  
[Publisher Full Text](#)
- Uliano-Silva M, Ferreira JGRN, Krashenninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- UniProt Consortium: **UniProt: a worldwide hub of protein knowledge.** *Nucleic Acids Res.* 2019; **47**(D1): D506–D515.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.  
[Publisher Full Text](#)
- 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 14 February 2025

<https://doi.org/10.21956/wellcomeopenres.25741.r117558>

© 2025 Kuraku S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



### Shigehiro Kuraku

National Institute of Genetics, Shizuoka, Japan

The manuscript soundly reports the chromosome-scale assembly of the common sole genome. I expect it to be recognized by researchers working on the species and used for fishery science as well as cross-species comparisons etc. for biodiversity studies in general. I suggest incorporating the changes requested below before it is accepted for indexing.

I wonder if the part 'Most of the assembly sequence (97.81%)' refers to the proportion in length or sequence number.

Is the sex of the animal used for sequencing unknown? If known, that should be clarified in the manuscript.

The details of the parameter choice in the Hi-C read mapping step (with bwa-mem2) should be included in the Methods. Also, wasn't the step to make duplicate reads (recommended by the YaHS developer) inserted?

#### 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:** vertebrate evolutionary genomics, developmental biology

**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 10 February 2025

<https://doi.org/10.21956/wellcomeopenres.25741.r117899>

© 2025 Marnis H. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Huria Marnis** 

National Research and Innovation Agency (BRIN), Cibinong, Indonesia

The present study is an interesting work providing genome sequence of common sole. It can be useful for further studies related to genome references in flatfish and other aquatic species. However, some minor issues exist in the text. They should be addressed before the manuscript is suitable for indexing.

#### **Genome sequence report**

Line 1 – use a common name to mention the species. The scientific name is only included at the first appearance of species in the text. Apply it thorough manuscript!

#### **Methods**

##### *Sample acquisition*

First paragraph, In the abstract, the authors mention that authors have chosen a single female specimen, but it should state in the methods and explain why a single female specimen was chosen for genome sequencing.

First paragraph, last sentence:

What were tissue samples collected?

*Was the specimen freshly caught or stored before processing?*

*Was it preserved immediately after collection to prevent degradation?*

Was it from a wild population or aquaculture setting? Mention it!

*Add details on the handling time from capture to preservation to ensure DNA/RNA integrity.*

First paragraph and second paragraph, The RNA-seq sample came from a different specimen (juvenile instead of adult female). Was this intentional? If so, how might developmental stage differences impact gene annotation?

##### *Nucleic acid extraction*

Mention DNA quality assessment (e.g., purity ratios, integrity checks).

##### *Library Preparation and Sequencing*

Were additional quality control steps performed?

**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:*** molecular genetic

**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.**

-----