DATA NOTE

The genome sequence of the European robin, *Erithacus rubecula* Linnaeus 1758 [version 1; peer review: awaiting peer review]

Jenny C. Dunn[1,2], Miriam Liedvogel[3,4], Michelle Smith[5], Craig Corton[5], Karen Oliver[5], Jason Skelton[5], Emma Betteridge[5], Jale Dolucan[5,6], Michael A. Quail[5], Marcela Uliano-Silva[5], Shane A. McCarthy[5,7], Kerstin Howe[5], James Torrance[5], Jonathan Wood[5], Sarah Pelan[5], Ying Sims[5], Richard Challis[5], Jonathan Threlfall[5], Daniel Mead[5,8], Mark Blaxter[5]

[1]Brayford Way, Brayford Pool, Lincoln, University of Lincoln, Lincoln, LN6 7TS, UK
[2]School of Biology, University of Leeds, Clarendon Way, Leeds, LS2 9JT, UK
[3]Max-Planck-Institut für Evolutionsbiologie, August-Thienemann-Str. 2, Plön, D-24306, Germany
[4]Institute of Avian Research, An der Vogelwarte 21, Wilhelmshaven, 26386, Germany
[5]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, UK
[6]Achilles Therapeutics plc, 245 Hammersmith Road, London, W6 8PW, UK
[7]Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
[8]Owlstone Medical, Cambridge Science Park, Cambridge, CB4 0GJ, UK

v1 **First published:** 02 Jul 2021, **6**:172
https://doi.org/10.12688/wellcomeopenres.16988.1
Latest published: 02 Jul 2021, **6**:172
https://doi.org/10.12688/wellcomeopenres.16988.1

Abstract

We present a genome assembly from an individual female *Erithacus rubecula* (the European robin; Chordata; Aves; Passeriformes; Turdidae). The genome sequence is 1.09 gigabases in span. The majority of the assembly is scaffolded into 36 chromosomal pseudomolecules, with both W and Z sex chromosomes assembled.

Keywords

Erithacus rubecula, European robin, genome sequence, chromosomal

Open Peer Review

Reviewer Status *AWAITING PEER REVIEW*

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

This article is included in the Tree of Life gateway.

Corresponding author: Mark Blaxter (mark.blaxter@sanger.ac.uk)

Author roles: Dunn JC: Data Curation, Formal Analysis, Investigation, Resources, Writing – Review & Editing; **Liedvogel M**: Data Curation, Formal Analysis, Investigation, Resources, Writing – Review & Editing; **Smith M**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Corton C**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Oliver K**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Skelton J**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Betteridge E**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Dolucan J**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Quail MA**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Uliano-Silva M**: Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **McCarthy SA**: Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **Howe K**: Formal Analysis, Investigation, Methodology, Software, Supervision, Validation, Writing – Review & Editing; **Torrance J**: Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **Wood J**: Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **Pelan S**: Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **Sims Y**: Formal Analysis, Investigation, Methodology, Software, Validation, Writing – Review & Editing; **Challis R**: Formal Analysis, Methodology, Software, Validation, Visualization, Writing – Review & Editing; **Threlfall J**: Project Administration, Writing – Original Draft Preparation, Writing – Review & Editing; **Mead D**: Conceptualization, Investigation, Project Administration, Writing – Review & Editing; **Blaxter M**: Conceptualization, Data Curation, Funding Acquisition, Supervision, Writing – Review & Editing

Competing interests: Jonathan Threlfall was employed by F1000 Research Limited until January 2021.

Grant information: This work was supported by the Wellcome Trust through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328). S.A.M. is supported by Wellcome (207492). M.L. is supported by SFB 1372 – Magnetoreception and Navigation in Vertebrates.
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: Dunn JC, Liedvogel M, Smith M *et al.* **The genome sequence of the European robin, *Erithacus rubecula* Linnaeus 1758 [version 1; peer review: awaiting peer review]** Wellcome Open Research 2021, **6**:172
https://doi.org/10.12688/wellcomeopenres.16988.1

First published: 02 Jul 2021, **6**:172 https://doi.org/10.12688/wellcomeopenres.16988.1

Species taxonomy

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Passeriformes; Turdidae; Erithacus; *Erithacus rubecula* Linnaeus 1758 (NCBI:txid37610).

Introduction

The European robin, *Erithacus rubecula*, is a small, insectivorous, partially migratory bird native to Europe, western Russia and Siberia, North Africa and the Middle East. Adults are predominantly brown with a characteristic red/orange breast; juveniles have spotted plumage and lack the red/orange breast. Robin populations are increasing both in the Atlantic archipelago of the United Kingdom (where it is the national bird) and Ireland, and worldwide (British Trust for Ornithology, 2019).

The robin is notable for being the first species in which the use of the earth's magnetic field for compass orientation during migration was described (Wiltschko & Wiltschko, 1972). The European robin continues to serve as an iconic model organism for migratory birds. Although the exact mechanism by which this magnetoreception occurs is not yet understood, two main complementary hypotheses are currently discussed. One is based on magnetite particles in the beak area of the bird and is mostly discussed in a map sense, and another hypothesis is based on a light-mediated biochemical reaction scheme (radical-pair reaction) that could mediate directional information provided by the earth's magnetic field into directional information for migratory journeys. The most promising receptor candidate for the latter light-mediated mechanism at current is cryptochrome 4, a blue light receptor molecule in the birds' eye (Günther *et al.*, 2018). The availability of a high quality annotated assembly of the robin's genome sequence will therefore enable researchers to investigate in more detail the genetic factors, such as cryptochrome 4, which drive robins to migrate and direct them where to go. As a model organism for behavioural research, the information deduced from the genetics of *E. rubecula* can then be used to understand the migratory behaviours of other bird species.

Genome sequence report

The reference genome was sequenced from one female *E. rubecula* collected from Eagle, Lincolnshire, UK. A total of 46-fold coverage in Pacific Biosciences single-molecule long reads (N50 19 kb) and 47-fold coverage in 10X Genomics read clouds (from molecules with an estimated N50 of 68 kb) were generated. Primary assembly contigs were scaffolded with chromosome conformation HiC data. Manual assembly curation corrected 110 missing/misjoins and removed 20 haplotypic duplications, reducing the scaffold number by 9.2%, increasing the scaffold N50 by 112.7% and decreasing the assembly length by 0.4%. The final assembly has a total length of 1.087 Gb in 1,120 sequence scaffolds with a scaffold N50 of 46.6 Mb (Table 1). The majority, 91.6%, of the assembly sequence was assigned to 36 chromosomal-level scaffolds representing 34 autosomes (numbered by sequence length), and the W and Z sex chromosomes (Figure 1–Figure 4; Table 2). The assembly has a BUSCO (Simão *et al.*, 2015) v5.0.0 completeness of 96.2% using the aves_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Table 1. Genome data for *Erithacus rubecula* bEriRub2.2.

Project accession data	
Assembly identifier	bEriRub2.2
Species	*Erithacus rubecula*
Specimen	bEriRub2
NCBI taxonomy ID	txid37610
BioProject	PRJEB38658
BioSample ID	SAMEA4760689
Isolate information	Female, blood
Raw data accessions	
PacificBiosciences SEQUEL I	ERX3338814, ERX3338816-ERX3338823
10X Genomics Illumina	ERX3341631-ERX3341634
Hi-C Illumina	ERX5308916
Genome assembly	
Assembly accession	GCA_903797595.2
Accession of alternate haplotype	GCA_903797565.1

Genome assembly	
Span (Mb)	1,087
Number of contigs	2,109
Contig N50 length (Mb)	5.59
Number of scaffolds	1120
Scaffold N50 length (Mb)	46.56
Longest scaffold (Mb)	112.1
BUSCO* genome score	C:96.2%[S:95.8%,D:0.4%],F:0.6%,M:3.2%,n:10,844

* BUSCO scores based on the aves_odb10 BUSCO set using v5.0.0. C= complete [S= single copy, D=duplicated], F=fragmented, M=missing, n=number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/Erithacus%20rubecula/dataset/CAIGJR02/busco.

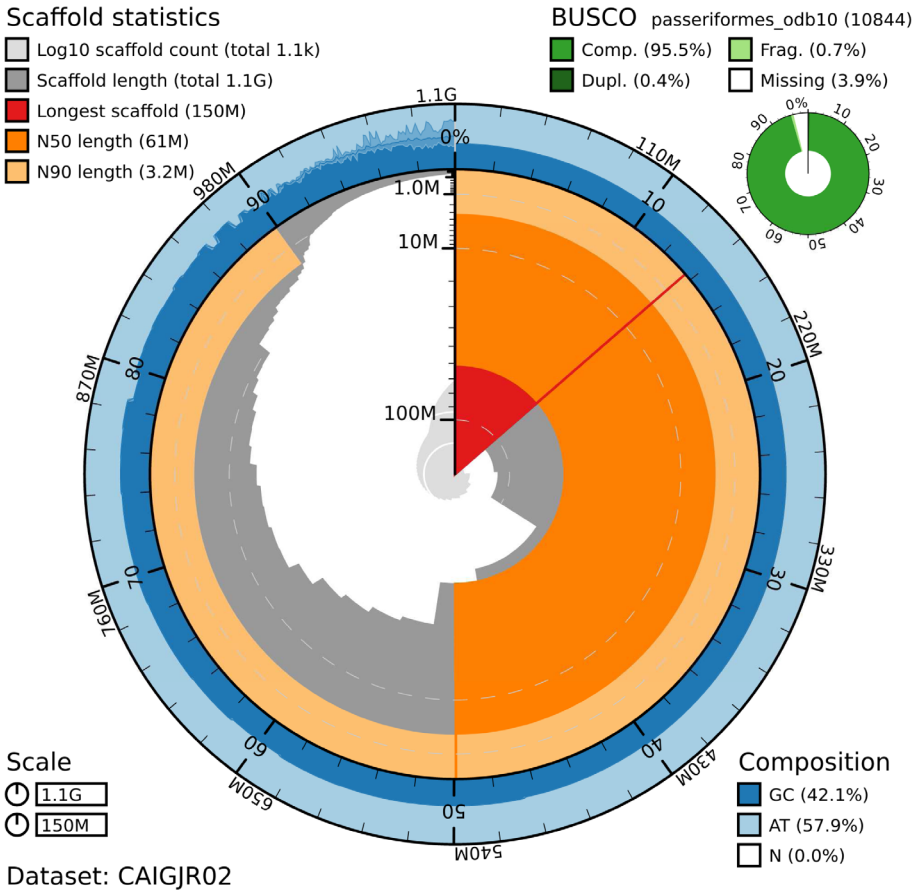

Figure 1. Genome assembly of *Erithacus rubecula*, bEriRub2.2. BlobToolKit Snailplot. The plot shows N50 metrics for bEriRub2 and BUSCO scores for the Passiformes set of orthologues. Interactive version available at https://blobtoolkit.genomehubs.org/view/Erithacus%20rubecula/dataset/CAIGJR02/snail.

Methods

A blood sample was taken from the brachael vein of a live bird during routine health checks of populations in Eagle, Lincolnshire, UK (latitude 53.193716, longitude -0.689135). Blood was collected through a glass capillary tube and stored at -20°C. The sample was taken under Home Office (ASPA) license number PB0AED9B7; birds were caught and handled under a British Trust for Ornithology ringing licence.

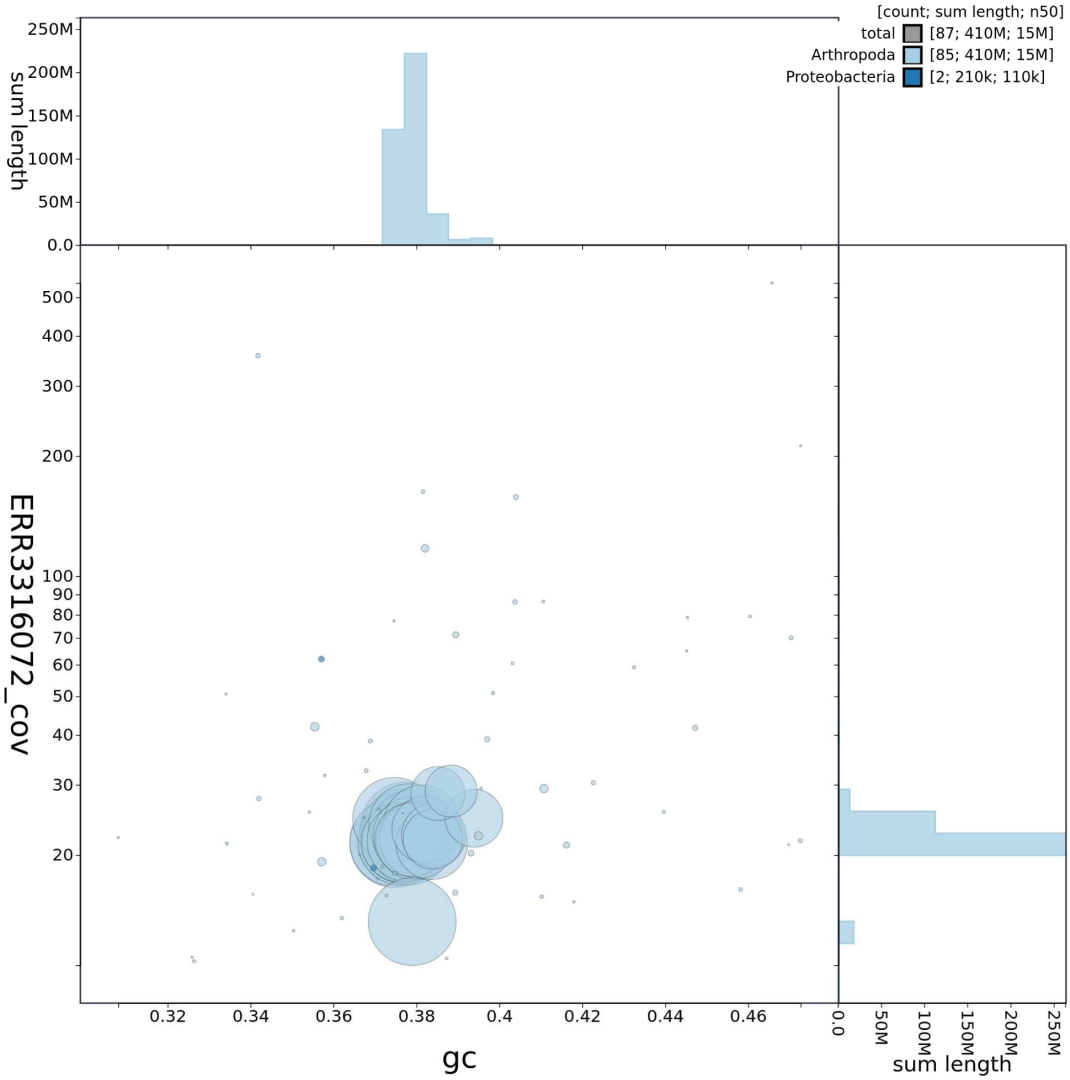

Figure 2. Genome assembly of *Erithacus rubecula*, bEriRub2.2. BlobToolKit GC-coverage plot. Interactive version available at https://blobtoolkit.genomehubs.org/view/Erithacus%20rubecula/dataset/CAIGJR02/blob?plotShape=circle.

Genomic DNA was extracted using an agarose plug extraction from a blood sample following the Bionano Prep Animal Tissue DNA Isolation Soft Tissue Protocol. Pacific Biosciences CLR long read and 10X Genomics read cloud sequencing libraries were constructed according to manufacturers' instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL I and Illumina HiSeq X instruments. Hi-C data were generated using the Dovetail HiC library preparation kit at the Wellcome Sanger Institute and sequenced using Illumina HiSeq X.

Assembly was carried out following the Vertebrate Genome Project pipeline v1.6 (Rhie *et al.*, 2020), without the use of Bionano data. Assembly was performed using Falcon-unzip (Chin *et al.*, 2016), haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2020) and a first round of scaffolding carried out with 10X Genomics read clouds using scaff10x. Scaffolding with Hi-C data (Rao *et al.*, 2014) was carried out with SALSA2 (Ghurye *et al.*, 2019). The Hi-C scaffolded assembly was polished with arrow using the PacBio data, then polished with the 10X Genomics Illumina data by aligning to the assembly with longranger align, calling variants with freebayes (Garrison & Marth, 2012) and applying homozygous non-reference edits using bcftools consensus. Two rounds of the Illumina polishing were applied. The assembly was checked for contamination and corrected using the gEVAL system (Chow *et al.*, 2016) as described previously (Howe *et al.*, 2021). Manual curation was performed using gEVAL, HiGlass and Pretext. Figure 1–Figure 3 and BUSCO scores were generated using BlobToolKit (Challis *et al.*, 2020). Software versions are given in Table 3.

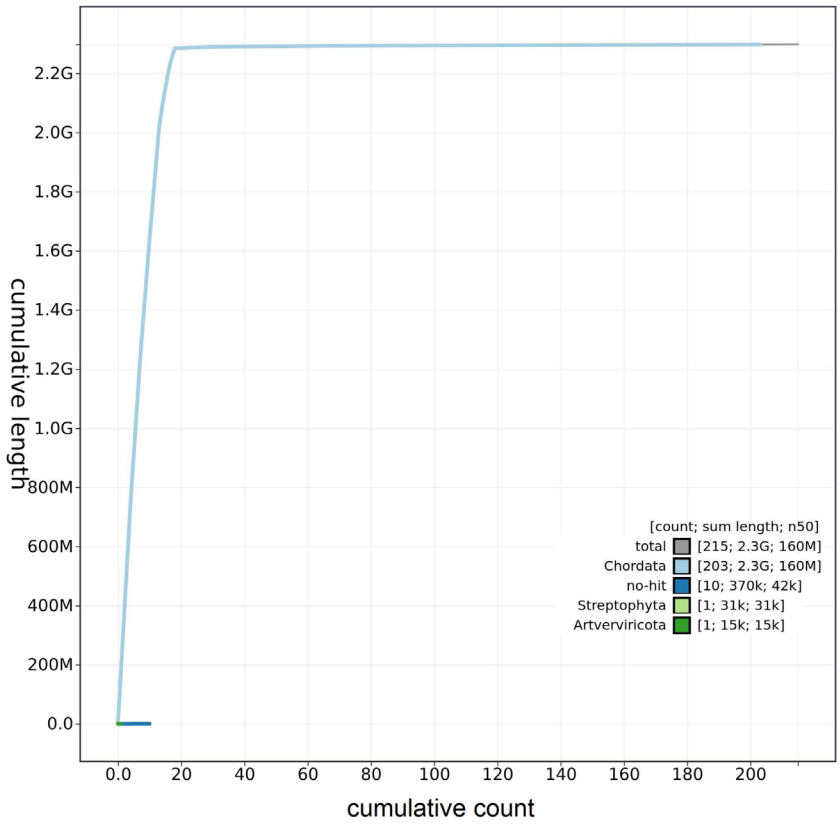

Figure 3. Genome assembly of *Erithacus rubecula*, bEriRub2.2: BlobToolKit Cumulative sequence plot. Interactive version available at https://blobtoolkit.genomehubs.org/view/Erithacus%20rubecula/dataset/CAIGJR02/cumulative.

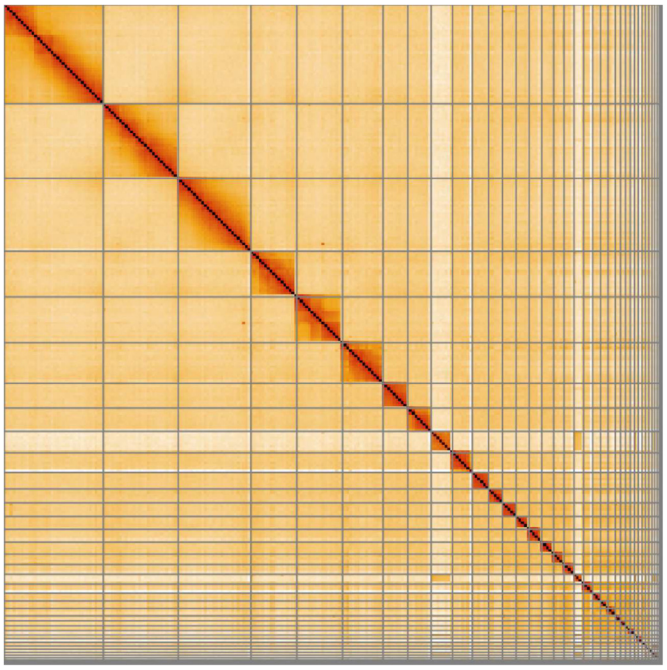

Figure 4. Genome assembly of *Erithacus rubecula*, bEriRub2.2: Hi-C contact map. Hi-C contact map of the bEriRub2.2 assembly, visualized in HiGlass (Kerpedjiev *et al.*, 2018).

Table 2. Chromosomal pseudomolecules in the genome assembly of *Erithacus rubecula* bEriRub2.2.

INSDC accession	Chromosome	Size (Mb)	GC%
LR812103.1	1	112.10	39.3
LR812104.1	2	109.05	39.7
LR812105.2	3	148.24	39.2
LR812106.1	4	68.60	39.9
LR812107.1	5	68.52	39.2
LR812108.1	6	60.68	41
LR812110.1	8	37.15	41.3
LR812111.1	9	34.93	41.8
LR812113.1	10	29.52	42.1
LR812114.1	11	24.63	43.1
LR812115.1	12	20.59	42.9
LR812116.1	13	20.40	44
LR812117.1	14	19.45	43.3
LR812118.1	15	19.11	43.5
LR812119.1	16	17.82	45
LR812120.1	17	15.59	45.4
LR812121.1	18	14.76	46.8
LR812122.1	19	13.42	46.4
LR812123.1	20	11.84	47.3
LR812124.1	21	11.12	48.5
LR812125.1	22	10.97	47.5
LR812126.1	23	7.61	50.5
LR812127.1	24	7.50	48.2
LR812128.1	25	7.13	49.1
LR812129.1	26	6.52	51.9
LR812131.1	27	5.43	52.2
LR812132.1	28	5.34	53
LR812133.1	29	4.77	50.2
LR812135.1	31	2.33	56.4
LR812130.2	W	4.15	44.6
LR812137.1	33	2.04	53.3
LR812112.1	Z	31.99	39.7
LR812138.1	34	0.96	49.3
	Unplaced	131.30	46

Table 3. Software tools used.

Software tool	Version	Source
Falcon-unzip	falcon-kit 1.2.2	(Chin *et al.*, 2016)
purge_dups	1.0.0	(Guan *et al.*, 2020)
scaff10x	4.2	https://github.com/wtsi-hpag/Scaff10X
arrow	GenomicConsensus 2.3.3	https://github.com/PacificBiosciences/GenomicConsensus
longranger align	2.2.2	https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines
freebayes	v1.1.0-3-g961e5f3	(Garrison & Marth, 2012)
bcftools consensus	1.9	http://samtools.github.io/bcftools/bcftools.html
gEVAL	2016	(Chow *et al.*, 2016)
HiGlass	1.11.6	(Kerpedjiev *et al.*, 2018)
PretextView	0.0.4	https://github.com/wtsi-hpag/PretextView
BlobToolKit	2.5	(Challis *et al.*, 2020)

Data availability

Underlying data

European Nucleotide Archive: Erithacus rubecula (European robin). Accession number PRJEB38659.

The genome sequence is released openly for reuse. The *E. rubecula* genome sequencing initiative is part of the Wellcome Sanger Institute's "25 genomes for 25 years" project. It is also part of the Vertebrate Genomes Project (VGP) ordinal references programme and the Darwin Tree of Life (DToL) project. All raw data and the assembly have been deposited in INSDC databases. The genome will be annotated and presented through the Ensembl pipeline at the European Bioinformatics Institute. Raw data and assembly accession identifiers are reported in Table 1.

Acknowledgements

We thank Mike Stratton and Julia Wilson for their support for the 25 genomes for 25 years project.

References

British Trust for Ornithology: **BirdTrends 2019: Trends in Numbers, Breeding Success and Survival for UK Breeding Birds**. 2019.
Reference Source

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit - Interactive Quality Assessment of Genome Assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–74.
PubMed Abstract | Publisher Full Text | Free Full Text

Chin CS, Peluso P, Sedlazeck FJ, *et al.*: **Phased Diploid Genome Assembly with Single-Molecule Real-Time Sequencing.** *Nat Methods.* 2016; **13**(12): 1050–54.
PubMed Abstract | Publisher Full Text | Free Full Text

Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL - a Web-Based Browser for Evaluating Genome Assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–10.
PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing.** arXiv: 1207.3907. 2012.
Reference Source

Ghurye RR, Sundaram K, Smith F, *et al.*: **Novel *ADA2* Mutation Presenting with Neutropenia, Lymphopenia and Bone Marrow Failure in Patients with Deficiency in Adenosine Deaminase 2 (DADA2).** *Br J Haematol.* 2019; **186**(3): e60–64.
PubMed Abstract | Publisher Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and Removing Haplotypic Duplication in Primary Genome Assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–98.
PubMed Abstract | Publisher Full Text | Free Full Text

Günther A, Einwich A, Sjulstok E, *et al.*: **Double-Cone Localization and Seasonal Expression Pattern Suggest a Role in Magnetoreception for European Robin Cryptochrome 4.** *Curr Biol.* 2018; **28**(2): 211–23.e4.
PubMed Abstract | 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

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

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): 1665–80.
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.** *bioRxiv.* 2020; 2020.05.22.110833.
Publisher Full Text

Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: Assessing Genome Assembly and Annotation Completeness with Single-Copy Orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–12.
PubMed Abstract | Publisher Full Text

Wiltschko W, Wiltschko R: **Magnetic Compass of European Robins.** *Science.* 1972; **176**(4030): 62–4.
PubMed Abstract | Publisher Full Text