DATA NOTE

The genome sequence of the European nightjar, *Caprimulgus europaeus* (Linnaeus, 1758) [version 1; peer review: 2 approved]

Simona Secomandi\(^1\), Fernando Spina\(^2\), Giulio Formenti\(^3,4\), Guido Roberto Gallo\(^1\), Manuela Caprioli\(^5\), Roberto Ambrosini\(^5\), Sara Riello\(^6\), Wellcome Sanger Institute Tree of Life programme, Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

\(^1\)Department of Biosciences, University of Milan, Milan, Italy
\(^2\)Institute for Environmental Protection and Research (ISPRA), Ozzano dell’Emilia, Italy
\(^3\)Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA
\(^4\)Howard Hughes Medical Institute, Chevy Chase, MD, USA
\(^5\)Department of Environmental Sciences and Policy, University of Milan, Milan, Italy
\(^6\)Riserva Naturale Statale “Isole di Ventotene e S. Stefano”, Ventotene, Italy

First published: 07 Dec 2021, 6:332
https://doi.org/10.12688/wellcomeopenres.17451.1

All rights reserved. This work is licensed under a Creative Commons Attribution 4.0 International License.

Open Peer Review

**Invited Reviewers**

1. Anne-Lyse Ducrest, University of Lausanne, Lausanne, Switzerland
2. Joshua Peñalba, Museum für Naturkunde, Berlin, Germany

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

Abstract

We present a genome assembly from an individual female *Caprimulgus europaeus* (the European nightjar; Chordata; Aves; Caprimulgiformes; Caprimulgidae). The genome sequence is 1,178 megabases in span. The majority of the assembly (99.33%) is scaffolded into 37 chromosomal pseudomolecules, including the W and Z sex chromosomes.

Keywords

Caprimulgus europaeus, European nightjar, Eurasian nightjar, genome sequence, chromosomal

This article is included in the Tree of Life gateway.
Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Secomandi S: Resources, Writing – Original Draft Preparation, Writing – Review & Editing; Spina F: Writing – Review & Editing; Formenti G: Conceptualization, Methodology, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing; Gallo GR: Resources, Writing – Review & Editing; Caprioli M: Writing – Review & Editing; Ambrosini R: Writing – Review & Editing; Riello S: Resources, Writing – Review & Editing;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (206194) and the Darwin Tree of Life Discretionary Award (218328). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2021 Secomandi S 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: Secomandi S, Spina F, Formenti G et al. The genome sequence of the European nightjar, *Caprimulgus europaeus* (Linnaeus, 1758) [version 1; peer review: 2 approved] Wellcome Open Research 2021, 6:332
https://doi.org/10.12688/wellcomeopenres.17451.1

First published: 07 Dec 2021, 6:332 https://doi.org/10.12688/wellcomeopenres.17451.1
Species taxonomy
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria; Aves; Neognathae; Caprimulgiformes; Caprimulgidae; Caprimulginae; Caprimulgus; *Caprimulgus europaeus* Linnaeus 1758 (NCBI: GCA_907165065.1).

Background
The European nightjar (*Caprimulgus europaeus*; also known as the Eurasian nightjar and common goatsucker) is an insectivorous, crepuscular, ground-nesting bird distributed throughout the Western Palearctic (Hagemeijer & Blair, 1997). It breeds in semi-natural dry and open habitats with scattered trees (Cram & Brooks, 1985). Little is known about the ecology of the European nightjar (Cram & Brooks, 1985; Polakowski et al., 2020), and in general that of the Caprimulgidae family. The family comprises peculiar species such as the only bird known to hibernate, the Common Poorwill (*Phalaenoptilus nuttallii*) (Carey, 2019; French, 2019; Woods et al., 2019), and one of the few birds that uses echo-localization, the South American Oilbird (*Steatornis caripensis*) (Brinklov et al., 2013). The European nightjar has been found to be more resistant to pathogens than other bird species (Jiang et al., 2021). Although categorized as ‘least concern’ by the IUCN (IUCN, 2016), the European nightjar has experienced a steady population decline in the past decades, and is of conservation concern in Europe (Eaton et al., 2015; Evens et al., 2017; Keller et al., 2010). The availability of a high-quality, chromosome-level reference genome will help to deepen the knowledge on the biology and evolution of this species, boosting studies on the genomics of the peculiar family of Caprimulgidae. Moreover, as genomic resources gain preeminence in conservation efforts (Allendorf, 2017; Fuentes-Pardo & Rizzante, 2017; Supple & Shapiro, 2018), we expect that the reference genome presented here will help aid planning conservation actions for the European nightjar.

Genome sequence report
The genome was sequenced from a blood sample taken from a single female *C. europaeus* collected from a bird ringing station in Ventotene, Italy (latitude 40.79404, longitude 13.42777). A total of 87-fold coverage in Pacific Biosciences single-molecule long reads and 62-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 144 missing/misjoins and removed 31 haplotypic duplications, reducing the assembly length by 0.15% and the scaffold number by 21.94%, and increasing the scaffold N50 by 26.46%.

The final assembly has a total length of 1,178 Mb in 121 sequence scaffolds with a scaffold N50 of 83 Mb (Table 1). Of the assembly sequence, 99.3% was assigned to 37 chromosomal-level scaffolds, representing 35 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) completeness of 97.4% (single 96.9%, duplicated 0.6%) using the aves_odb10 reference set. While not fully phased, the assembly deposited is of one pseudo-haplotype. Contigs corresponding to the alternate haplotype have also been deposited.

### Methods

#### Sample acquisition
Sampling was performed during the routine activity of the scientific ringing station located in Ventotene island, Latina, Italy (latitude 40.7926⁰, longitude 13.4241⁰) during spring migration. Samples have been collected by ISPRa researchers within their institutional activities as from Italian national Law n. 157/92. Bird capture was performed in the evening according to standardized protocols using mist-nets (Saino et al., 2010; Spina et al., 1993). The sample was collected with a heparinized capillary tube after puncturing the ulnar

#### Table 1. Genome data for *Caprimulgus europaeus*, bCapEur3.1.

<table>
<thead>
<tr>
<th>Project accession data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly identifier</td>
</tr>
<tr>
<td>Species</td>
</tr>
<tr>
<td>Specimen</td>
</tr>
<tr>
<td>NCBI taxonomy ID</td>
</tr>
<tr>
<td>BioProject</td>
</tr>
<tr>
<td>BioSample ID</td>
</tr>
<tr>
<td>Isolate information</td>
</tr>
<tr>
<td>Raw data accessions</td>
</tr>
<tr>
<td>PacificBiosciences SEQUEL II</td>
</tr>
<tr>
<td>10X Genomics Illumina</td>
</tr>
<tr>
<td>Hi-C Illumina</td>
</tr>
<tr>
<td>Genome assembly</td>
</tr>
<tr>
<td>Assembly accession</td>
</tr>
<tr>
<td>Accession of alternate haplotype</td>
</tr>
<tr>
<td>Span (Mb)</td>
</tr>
<tr>
<td>Number of contigs</td>
</tr>
<tr>
<td>Contig N50 length (Mb)</td>
</tr>
<tr>
<td>Number of scaffolds</td>
</tr>
<tr>
<td>Scaffold N50 length (Mb)</td>
</tr>
<tr>
<td>Longest scaffold (Mb)</td>
</tr>
<tr>
<td>BUSCO* genome score</td>
</tr>
</tbody>
</table>

*BUSCO scores based on the aves_odb10 BUSCO set using v5.1.2. C= complete [S= single copy, D=duplicated], F=fragmented, M=missing, n=number of orthologues in comparison. A full set of BUSCO scores is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAvO11/buso.

Figure 1
Figure 1. Genome assembly of *Caprimulgus europaeus*, bCapEur3.1: metrics. The BlobToolKit Snailplot shows N50 metrics and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 1,177,791,212 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (126,318,510 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (82,614,289 and 15,699,869 bp), respectively. The pale grey spiral shows the cumulative chromosome 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 aves_odb10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/snail.

DNA extraction and sequencing

High molecular weight DNA was extracted from the blood sample at the Scientific Operations core of the Wellcome Sanger Institute using the Bionano Prep Blood DNA Isolation Kit according to the Bionano Prep Frozen Blood protocol. Pacific Biosciences CLR long read and 10X Genomics read cloud sequencing libraries were constructed according to the manufacturers’ instructions. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II and Illumina HiSeq X instruments. Hi-C data were generated from the same blood sample using the Arima Hi-C+ kit and sequenced on HiSeq X.

Genome assembly

Assembly was carried out following the Vertebrate Genome Project pipeline v1.6 (Rhie et al., 2020) with Falcon-unzip (Chin et al., 2016), haplotypic duplication was identified and
Figure 2. Genome assembly of Caprimulgus europaeus, bCapEur3.1: GC coverage. BlobToolKit GC-coverage plot. Scaffolds are coloured by phylum. Circles are sized in proportion to scaffold length. Histograms show the distribution of scaffold length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/blob.

removed with purge_dups (Guan et al., 2020) and a first round of scaffolding carried out with 10X Genomics read clouds using scaf10x. 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, with merfin (Formenti et al., 2021b) applied to avoid a drop in QV, 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. A complete mitochondrion was not found using mitoVGP (Formenti et al., 2021a), likely due to the sample being sourced from blood tissue, so mitochondrial sequence NC_025773.1 (Caprimulgus indicus) was used during
Figure 3. Genome assembly of *Caprimulgus europaeus*, bCapEur3.1: cumulative sequence. BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/bCapEur3.1/dataset/CAJRAV01/cumulative.
polishing. 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 (Howe et al., 2021) was performed using gEVAL, HiGlass (Kerpedjiev et al., 2018) and Pretex. The genome was analysed, and BUSCO scores generated, within the BlobToolKit environment (Challis et al., 2020). Table 3 gives version numbers of the software tools used in this work.
Table 2. Chromosomal pseudomolecules in the genome assembly of Caprimulgus europaeus, bCapEur3.1.

<table>
<thead>
<tr>
<th>INSDC accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>OU015523.1</td>
<td>1</td>
<td>126.32</td>
<td>40.1</td>
</tr>
<tr>
<td>OU015524.1</td>
<td>2</td>
<td>125.37</td>
<td>40.3</td>
</tr>
<tr>
<td>OU015525.1</td>
<td>3</td>
<td>100.16</td>
<td>39.8</td>
</tr>
<tr>
<td>OU015526.1</td>
<td>4</td>
<td>83.32</td>
<td>39.9</td>
</tr>
<tr>
<td>OU015528.1</td>
<td>5</td>
<td>82.61</td>
<td>40.7</td>
</tr>
<tr>
<td>OU015529.1</td>
<td>6</td>
<td>65.35</td>
<td>41.7</td>
</tr>
<tr>
<td>OU015530.1</td>
<td>7</td>
<td>60.47</td>
<td>40.6</td>
</tr>
<tr>
<td>OU015531.1</td>
<td>8</td>
<td>50.91</td>
<td>42.8</td>
</tr>
<tr>
<td>OU015532.1</td>
<td>9</td>
<td>48.66</td>
<td>41.6</td>
</tr>
<tr>
<td>OU015533.1</td>
<td>10</td>
<td>43.00</td>
<td>41.3</td>
</tr>
<tr>
<td>OU015534.1</td>
<td>11</td>
<td>35.23</td>
<td>42.1</td>
</tr>
<tr>
<td>OU015535.1</td>
<td>12</td>
<td>23.52</td>
<td>43.4</td>
</tr>
<tr>
<td>OU015536.1</td>
<td>13</td>
<td>22.81</td>
<td>42.3</td>
</tr>
<tr>
<td>OU015538.1</td>
<td>14</td>
<td>22.35</td>
<td>43.3</td>
</tr>
<tr>
<td>OU015539.1</td>
<td>15</td>
<td>19.40</td>
<td>42.8</td>
</tr>
<tr>
<td>OU015540.1</td>
<td>16</td>
<td>18.74</td>
<td>45</td>
</tr>
<tr>
<td>OU015541.1</td>
<td>17</td>
<td>16.93</td>
<td>45.6</td>
</tr>
<tr>
<td>OU015542.1</td>
<td>18</td>
<td>15.70</td>
<td>45.4</td>
</tr>
</tbody>
</table>

Table 3. Software tools used.

<table>
<thead>
<tr>
<th>Software tool</th>
<th>Version</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falcon-unzip</td>
<td>1.8.0</td>
<td>Chin et al., 2016</td>
</tr>
<tr>
<td>purge_dups</td>
<td>1.2.3</td>
<td>Guan et al., 2020</td>
</tr>
<tr>
<td>SALSA2</td>
<td>2.2</td>
<td>Ghurye et al., 2019</td>
</tr>
<tr>
<td>Arrow</td>
<td>GCpp-1.9.0</td>
<td><a href="https://github.com/PacificBiosciences/GenomicConsensu">https://github.com/PacificBiosciences/GenomicConsensu</a></td>
</tr>
<tr>
<td>Merfin</td>
<td>1.7</td>
<td>Formenti et al., 2021b</td>
</tr>
<tr>
<td>longranger align</td>
<td>2.2.2</td>
<td><a href="https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines">https://support.10xgenomics.com/genome-exome/software/pipelines/latest/advanced/other-pipelines</a></td>
</tr>
<tr>
<td>freebayes</td>
<td>1.3.1-17-gaa2ace8</td>
<td>Garrison &amp; Marth, 2012</td>
</tr>
<tr>
<td>gEVAL</td>
<td>N/A</td>
<td>Chow et al., 2016</td>
</tr>
<tr>
<td>HiGlass</td>
<td>1.11.6</td>
<td>Kerpedijev et al., 2018</td>
</tr>
<tr>
<td>PretextView</td>
<td>0.1.x</td>
<td><a href="https://github.com/wtsi-hpag/PretextView">https://github.com/wtsi-hpag/PretextView</a></td>
</tr>
<tr>
<td>BlobToolKit</td>
<td>2.6.2</td>
<td>Challis et al., 2020</td>
</tr>
</tbody>
</table>
Data availability

The genome sequence is released openly for reuse. The C. europaeus genome sequencing initiative is part of the Darwin Tree of Life (DToL) project and the Vertebrate Genomes Project. All raw sequence 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.

References


Carey C: Life In The Cold: Ecological, Physiological, and Molecular Mechanisms. CRC Press, 2019. Publisher Full Text


Publisher Full Text


Author information
Members of the Darwin Tree of Life Consortium are listed here: https://doi.org/10.5281/zenodo.4783559.

Members of the Darwin Tree of Life Barcoding collective are listed here: https://doi.org/10.5281/zenodo.4893704.

Members of the Wellcome Sanger Institute Tree of Life collective are listed here: https://doi.org/10.5281/zenodo.4783586.


Members of the Tree of Life Core Informatics collective are listed here: https://doi.org/10.5281/zenodo.5013542.
Open Peer Review

Current Peer Review Status: ✔️ ✔️

Version 1

Reviewer Report 04 January 2022

https://doi.org/10.21956/wellcomeopenres.19297.r47480

© 2022 Peñalba J. 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.

Joshua Peñalba
Center for Integrative Biodiversity Discovery, Museum für Naturkunde, Berlin, Germany

The authors describe the sequencing and assembly of the chromosome-scale reference genome for the European Nightjar. The methods follow that of the Vertebrate Genome Project pipeline. I just have some minor comments:

- How was the bird identified as female?
- About how much blood was used for the sequencing?
- How was the quality of the DNA checked?
- How many PacBio cells and Illumina lanes were used for each sequencing method?
- How did you know how many chromosomes should have been assembled?
- Can you provide more details on the assembly, which parameters were used and how was manual curation performed? If this is detailed in a different manuscript, please explicitly state which manuscript.

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?
Partly

Are the datasets clearly presented in a useable and accessible format?
Yes
Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomics, evolution, population genomics

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.

Anne-Lyse Ducrest
Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

The authors described a nice almost complete genome with pseudo-chromosomes of the European nightjar using PacBio Sequel II, Illumina, and HiCi sequencing methods and thus present important data for further genetic analysis.

There are two points that could be improved:
- There are some redundancies between Figures 1, 2, and Table 1.
- The method how to get long HMV DNA is not well described since the Bionano protocol is for human blood and not for bird blood.

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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: genomic, molecular 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.