DATA NOTE

The genome sequence of the Eurasian red squirrel, *Sciurus vulgaris* Linnaeus 1758 [version 1; peer review: 2 approved]

Daniel Mead, Kathryn Fingland, Rachel Cripps, Roberto Portela Miguez, Michelle Smith, Craig Corton, Karen Oliver, Jason Skelton, Emma Betteridge, Jale Duluocan, Olga Dudchenko, Arina D. Omer, David Weisz, Erez Lieberman Aiden, Olivier Fedrigo, Jacquelyn Mountcastle, Erich Jarvis, Shane A. McCarthy, Ying Sims, James Torrance, Alan Tracey, Kerstin Howe, Richard Challis, Richard Durbin, Mark Blaxter

1Tree of Life, Wellcome Sanger Institute, Cambridge, CB10 1SA, UK
2School of Animal, Rural and Environmental Sciences, Nottingham Trent University, Nottingham, NG25 0QF, UK
3The Wildlife Trust for Lancashire, Manchester and North Merseyside, Preston, PR5 6BY, UK
4Department of Life Sciences, Natural History Museum, London, SW7 5BD, UK
5Baylor College of Medicine, Houston, TX, 77030, USA
6Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, 10065, USA
7Howard Hughes Medical Institute, Chevy Chase, MD, 20815, USA
8Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK

First published: 03 Feb 2020, 5:18
https://doi.org/10.12688/wellcomeopenres.15679.1

Abstract

We present a genome assembly from an individual male *Sciurus vulgaris* (the Eurasian red squirrel; Vertebrata; Mammalia; Eutheria; Rodentia; Sciuridae). The genome sequence is 2.88 gigabases in span. The majority of the assembly is scaffolded into 21 chromosomal-level scaffolds, with both X and Y sex chromosomes assembled.

Keywords

*Sciurus vulgaris*, red squirrel, genome sequence, chromosomal

This article is included in the Tree of Life gateway.

Open Peer Review

Reviewer Status

Invited Reviewers

1

2

version 1

03 Feb 2020

report

report

1. Peter H. Sudmant, University of California, Berkeley, Berkeley, USA

2. Rob Ogden, University of Edinburgh, Edinburgh, UK

Any reports and responses or comments on the article can be found at the end of the article.
Corresponding author: Mark Blaxter (mark.blaxter@sanger.ac.uk)

Author roles: Mead D: Conceptualization, Investigation, Writing – Review & Editing; Fingland K: Investigation, Resources, Writing – Review & Editing; Cripps R: Investigation, Resources, Writing – Review & Editing; Portela Miguez R: Investigation, Resources, Writing – Review & Editing; Smith M: Investigation, Methodology, Writing – Review & Editing; Corton C: Investigation, Methodology, Writing – Review & Editing; Oliver K: Methodology, Supervision, Writing – Review & Editing; Skelton J: Methodology, Writing – Review & Editing; Betteridge E: Methodology, Writing – Review & Editing; Dolucan J: Methodology, Software, Writing – Review & Editing; Dudchenko O: Investigation, Methodology, Software, Visualization, Writing – Review & Editing; Weiss D: Methodology, Software, Visualization, Writing – Review & Editing; Lieberman Aiden E: Funding Acquisition, Supervision, Writing – Review & Editing; Fedrigo O: Investigation, Methodology, Writing – Review & Editing; McCarthy SA: Investigation, Methodology, Software, Supervision, Writing – Review & Editing; Sims Y: Investigation, Software, Visualization, Writing – Review & Editing; Mountcastle J: Data Curation, Writing – Review & Editing; Jarvis E: Data Curation, Writing – Review & Editing; Tracey A: Data Curation, Writing – Review & Editing; Torrance J: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; Challis R: Conceptualization, Funding Acquisition, Supervision, Writing – Review & Editing; Blaxter M: Conceptualization, Data Curation, Funding Acquisition, Project Administration, Resources, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by the Wellcome Trust through core funding to the Wellcome Sanger Institute (206194). SMcC and RD were supported by Wellcome grant 207492. ELA was supported by an NSF Physics Frontiers Center Award (PHY1427654), the Welch Foundation (Q-1866), a USDA Agriculture and Food Research Initiative Grant (2017-05741), and an NIH Encyclopedia of DNA Elements Mapping Center Award (UM1HG009375). RC was supported by BBSRC grant BB/R015325/2. The Darwin Tree of Life project is supported by Wellcome grant 218328.

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

Copyright: © 2020 Mead D 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: Mead D, Fingland K, Cripps R et al. The genome sequence of the Eurasian red squirrel, Sciurus vulgaris Linnaeus 1758 [version 1; peer review: 2 approved] Wellcome Open Research 2020, 5:18 https://doi.org/10.12688/wellcomeopenres.15679.1

First published: 03 Feb 2020, 5:18 https://doi.org/10.12688/wellcomeopenres.15679.1
Species taxonomy
Eukaryota; Metazoa; Chordata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciuromorpha; Sciuridae; Sciurinae; Sciurini; Sciurus; Sciurus vulgaris Linnaeus 1758 (NCBI txid 55149).

Background
The Eurasian red squirrel, Sciurus vulgaris, is native to northern Eurasia. In the Atlantic Archipelago of Britain and Ireland, S. vulgaris is under threat from anthropogenic pressure on its native woodland habitats1, and from competition from the introduced American grey squirrel, Sciurus carolinensis, particularly mediated by squirrelpox virus (Chantrey et al., 2014). The current population of S. vulgaris in the Atlantic Archipelago is estimated to be 150,000, and there are extensive efforts to conserve this species and expand its range (Hardouin et al., 2019). Here we present a chromosomally assembled genome sequence for S. vulgaris, based on a male specimen from Britain. This genome sequence will be of utility in population genomic analysis of fragmented S. vulgaris populations (Barratt et al., 1999), in managing reintroductions and in investigating the biology of susceptibility to squirrelpox virus (Darby et al., 2014).

Genome sequence report
The genome was sequenced from DNA extracted from a naturally deceased male S. vulgaris collected as part of a squirrel monitoring project run by the Wildlife Trust for Lancashire, Manchester and North Merseyside. A total of 51-fold coverage in Pacific Biosciences single-molecule long reads (N50 19 kb) and 44-fold coverage in 10X Genomics read clouds (from molecules with an estimated N50 of 69 kb) were generated. Primary assembly contigs were scaffolded with 10X reads, chromosome conformation HiC data, and 111-fold coverage of Bionano optical maps. The final assembly has a total length of 2.88 Gb in 638 sequence scaffolds with a scaffold N50 of 153.9 Mb (Table 1). The majority, 92.7%, of the

<table>
<thead>
<tr>
<th>Table 1. Genome data for Sciurus vulgaris mSciVul1.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project accession data</strong></td>
</tr>
<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><strong>Raw data accessions</strong></td>
</tr>
<tr>
<td>PacificBiosciences SEQUEL I</td>
</tr>
<tr>
<td>10X Genomics Illumina</td>
</tr>
<tr>
<td>Hi-C Illumina</td>
</tr>
<tr>
<td>BioNano data and assembly</td>
</tr>
<tr>
<td><strong>Genome assembly</strong></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>

assembly sequence was assigned to 21 chromosomal pseu-
domolecules representing 19 autosomes (numbered by sequence
length), and the X and Y sex chromosomes (Figure 1–Figure 4;
Table 2). The assembly has a BUSCO (Simão et al., 2015) com-
pleteness of 93.8% using the mammalia_odb9 reference set. The
primary assembly is a large-scale mosaic of both haplotypes
(i.e. is not fully phased) and we have therefore also deposited
the contigs corresponding to the alternate haplotype. The
genome can be compared to that of the grey squirrel, Sciurus
carolinensis, which we have also assembled.

Methods
The red squirrel specimen was collected from a garden in
Beechwood Drive, Formby, Merseyside, L37 2DQ. Grid ref:
SD2829706400 (Lat Long: 53.549316, -3.0836773) by the Wildlife
Trust for Lancashire, Manchester and North Merseyside as part
of an ongoing programme of recovery of dead squirrels. The
spleen was dissected out during autopsy. A full tissue dissec-
tion and preservation in 80% ethanol was undertaken and the
specimen accessioned by the Natural History Museum, London.

DNA was extracted using an agarose plug extraction from spleen
tissue following the Bionano Prep Animal Tissue DNA Isola-
tion Soft Tissue Protocol2. Pacific Biosciences CLR long read
and 10X Genomics read cloud sequencing libraries were con-
structed according to the 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 by the Aiden

2 https://bionanogenomics.com/wp-content/uploads/2018/02/30077-Bio-

Figure 1. Genome assembly of Sciurus vulgaris mSciVul1: Metrics. BlobToolKit Snailplot showing N50 metrics for S. vulgaris assembly
mSciVul1 and BUSCO scores for the Euarchontoglires set of orthologues. The interactive version of this figure is available here.

Page 4 of 12
Figure 2. Genome assembly of Sciurus vulgaris mSciVul1: GC-coverage plot. BlobToolKit GC-coverage plot of S. vulgaris mSciVul1. The interactive version of this figure is available here.

Lab using an optimised version of their protocols (Dudchenko et al., 2017). BioNano data were generated in the Rockefeller University Vertebrate Genome laboratory using the Saphyr instrument. Ultra-high molecular weight DNA was extracted using the Bionano Prep Animal Tissue DNA Isolation Soft Tissue Protocol and assessed by pulsed field gel and Qubit 2 fluorimetry. DNA was labeled for Bionano Genomics optical mapping following the Bionano Prep Direct Label and Stain (DLS) Protocol and run on one Saphyr instrument chip flowcell. The total yield of tagged molecules ≥ 150 kb with at least 9
Figure 3. Genome assembly of *Sciurus vulgaris* mSciVul1: Cumulative sequence plot. Dashed line shows the cumulative sequence plot of *S. carolinensis* mSciCar1 for comparison. The interactive version of this figure is available here.

sites was 320.6 Gb (N50 0.25 Mb). A CMAP (Bionano assembly consensus genome map) was de-novo assembled using Bionano Solve (see Table 3 for software versions and sources) yielding 574 maps with a total map length of 3.28 Gb and a map N50 of 86.34 Mb.

Assembly followed a modified version of the Vertebrate Genomes Project assembly protocols\(^1\). In brief, assembly was carried

\(^1\)https://github.com/VGP/vgp-tools
Figure 4. Genome assembly of *Sciurus vulgaris* mSciVul1: Hi-C contact map. Hi-C contact map of the *S. vulgaris* mSciVul1 assembly, visualized in Juicebox. The interactive version of this figure is available [here](#), powered by Juicebox.js (Robinson *et al.*, 2018).

Out using Falcon-unzip (Chin *et al.*, 2016), haplotypic duplication was identified and removed with purge_dups (Guan *et al.*, 2019) and a first round of scaffolding carried out with 10X Genomics read clouds using scaff10x. Hybrid scaffolding was performed using the BioNano DLE-1 data and BioNano Solve.

Scaffolding with Hi-C data (Rao *et al.*, 2014) was carried out with 3D-DNA (Dudchenko *et al.*, 2017), followed by manual curation with Juicebox Assembly Tools (Dudchenko *et al.*, 2018; Durand *et al.*, 2016; Robinson *et al.*, 2018). The Hi-C scaffolded assembly was polished using arrow with 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
**Table 2.** Chromosomal pseudomolecules in the genome assembly of *Sciurus vulgaris* mSciVu1.

<table>
<thead>
<tr>
<th>ENA accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR738611.1</td>
<td>1</td>
<td>213.19</td>
<td>40.3</td>
</tr>
<tr>
<td>LR738612.1</td>
<td>2</td>
<td>204.37</td>
<td>40.9</td>
</tr>
<tr>
<td>LR738613.1</td>
<td>3</td>
<td>189.66</td>
<td>40.4</td>
</tr>
<tr>
<td>LR738614.1</td>
<td>4</td>
<td>187.75</td>
<td>39.4</td>
</tr>
<tr>
<td>LR738615.1</td>
<td>5</td>
<td>181.67</td>
<td>39.6</td>
</tr>
<tr>
<td>LR738616.1</td>
<td>6</td>
<td>173.03</td>
<td>39</td>
</tr>
<tr>
<td>LR738617.1</td>
<td>7</td>
<td>162.82</td>
<td>39.4</td>
</tr>
<tr>
<td>LR738618.1</td>
<td>8</td>
<td>153.87</td>
<td>40.6</td>
</tr>
<tr>
<td>LR738619.1</td>
<td>9</td>
<td>146.52</td>
<td>38.2</td>
</tr>
<tr>
<td>LR738620.1</td>
<td>10</td>
<td>145.29</td>
<td>38.9</td>
</tr>
<tr>
<td>LR738622.1</td>
<td>11</td>
<td>132.26</td>
<td>40.2</td>
</tr>
<tr>
<td>LR738623.1</td>
<td>12</td>
<td>115.30</td>
<td>40.2</td>
</tr>
<tr>
<td>LR738624.1</td>
<td>13</td>
<td>100.60</td>
<td>40.8</td>
</tr>
<tr>
<td>LR738625.1</td>
<td>14</td>
<td>99.24</td>
<td>41.3</td>
</tr>
<tr>
<td>LR738626.1</td>
<td>15</td>
<td>91.00</td>
<td>40.2</td>
</tr>
<tr>
<td>LR738627.1</td>
<td>16</td>
<td>79.70</td>
<td>42.9</td>
</tr>
<tr>
<td>LR738628.1</td>
<td>17</td>
<td>75.03</td>
<td>45.1</td>
</tr>
<tr>
<td>LR738629.1</td>
<td>18</td>
<td>41.58</td>
<td>47.9</td>
</tr>
<tr>
<td>LR738630.1</td>
<td>19</td>
<td>33.12</td>
<td>45.1</td>
</tr>
<tr>
<td>LR738621.1</td>
<td>X</td>
<td>138.34</td>
<td>37.9</td>
</tr>
<tr>
<td>LR738631.1</td>
<td>Y</td>
<td>4.04</td>
<td>38.9</td>
</tr>
<tr>
<td>-</td>
<td>Unplaced</td>
<td>210.22</td>
<td>15.6</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>BioNano Solve</td>
<td>3.3</td>
<td><a href="http://www.bnxinstall.com/solve/BionanoSolveInstall.html">http://www.bnxinstall.com/solve/BionanoSolveInstall.html</a></td>
</tr>
<tr>
<td>Falcon-unzip</td>
<td>falcon-kit 1.1.1</td>
<td>Chin et al., 2016</td>
</tr>
<tr>
<td>purge_dups</td>
<td>1.0.0</td>
<td>Guan et al., 2019</td>
</tr>
<tr>
<td>scaff10x</td>
<td>4.2</td>
<td><a href="https://github.com/wtsi-hpag/Scaff10X">https://github.com/wtsi-hpag/Scaff10X</a></td>
</tr>
<tr>
<td>3D-DNA</td>
<td>180419</td>
<td>Dudchenko et al., 2017</td>
</tr>
<tr>
<td>Juicebox Assembly Tools</td>
<td>1.9.8</td>
<td>Dudchenko et al., 2018</td>
</tr>
<tr>
<td>arrow</td>
<td>GenomicConsensus 2.3.3</td>
<td><a href="https://github.com/PacificBiosciences/GenomicConsensus">https://github.com/PacificBiosciences/GenomicConsensus</a></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>v1.1.0-3-g961e5f3</td>
<td>Garrison &amp; Marth, 2012</td>
</tr>
<tr>
<td>bcftools consensus</td>
<td>1.9</td>
<td><a href="http://samtools.github.io/bcftools/bcftools.html">http://samtools.github.io/bcftools/bcftools.html</a></td>
</tr>
<tr>
<td>gEVAL</td>
<td>2016</td>
<td>Chow et al., 2016</td>
</tr>
<tr>
<td>BlobToolKit</td>
<td>1</td>
<td>Challis et al., 2019</td>
</tr>
</tbody>
</table>
checked for contamination and further manually assessed and corrected using the gEVAL system (Chow et al., 2016). The genome was analysed within the BlobToolKit environment (Challis et al., 2019).

Data availability

Underlying data

The genome sequence is released openly for reuse. The S. vulgaris genome sequencing initiative is part of the Wellcome Sanger Institute’s “25 genomes for 25 years” project 1. It is also part of the Vertebrate Genomes Project (VGP) 4 ordinal references programme, the DNA Zoo Project 6 and the Darwin Tree of Life (DToL) project 7. The specimen has been preserved in ethanol and deposited with the Natural History Museum, London under registration number NHMUK ZD 2019.213, where it will remain accessible to the research community for posterity. All raw sequence data and the assembly have been deposited in the ENA. 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.

Author contributions

Collection and identification: KF, RC

DNA extraction and sequencing: DM, MS, CC, KO, JS, EB, JD, ADO, OF, JM, EJ

Genome assembly and curation: SMcC, OD, DW, ELA, KH, RC, YS, JT, AT

Project management: DM, RD, MB

Manuscript: MB, assisted by all authors

Acknowledgements

We thank Mike Stratton and Julia Wilson for their continuing support for the 25 genomes for 25 years project. The Wildlife Trust for Lancashire, Manchester and North Merseyside thank many members of the public for support.

References


Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 19 June 2020

https://doi.org/10.21956/wellcomeopenres.17184.r38866

© 2020 Ogden R. 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.

Rob Ogden
Royal (Dick) School of Veterinary Studies and the Roslin Institute, University of Edinburgh, Edinburgh, UK

The article details the production and release of the first genome sequence for the Eurasian red squirrel, *Sciurus vulgaris*, generated from an individual squirrel sample collected in the United Kingdom. It provides a comprehensive explanation of the sequencing and assembly methods used, and resulting data availability, alongside a summary of key genome assembly characteristics. The release of the genome will provide an important reference sequence resource for future studies of red squirrel biology, in particular, investigations of immunogenetic diversity and population genetic research to support conservation management.

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

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

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

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

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

**Reviewer Expertise:** Evolutionary and population genetics, applied conservation genetics.

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.
Peter H. Sudmant

Department of Integrative Biology, University of California, Berkeley, Berkeley, CA, USA

In this manuscript Mead and colleagues report on a draft assembly of the Eurasian red squirrel, *Sciurus vulgaris*. The genome assembly is constructed from a combination of Pac-bio and 10X, HiC, and Bionano optical mapping, using a slightly modified standard VGP assembly protocols. The resulting assembly is largely assigned to 21 chromosomes with high scaffold N50 (~150 Mb). Overall the manuscript concisely describes the genome and the presented resources will be of great use for population genetics on this species and related taxa. Some additional minor details on the assembly quality would be informative.

Specific comments:
- Mention of the contig N50 would be helpful in addition to the number of gaps.
- I cannot find any information about the base-pair quality of this reference? That too would be helpful.
- Citing the accompanying grey squirrel manuscript would be useful, particularly as this genome is referred to in Figure 3.

References

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: Genetics and 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.