DATA NOTE

The genome sequence of the eastern grey squirrel, *Sciurus carolinensis* Gmelin, 1788 [version 1; peer review: 2 approved]

Dan Mead, Kathryn Fingland, Rachel Cripps, Roberto Portela Miguez, Michelle Smith, Craig Corton, Karen Oliver, Jason Skelton, Emma Betteridge, Jale Doulcan, Michael A. Quail, Shane A. McCarthy, Kerstin Howe, Ying Sims, James Torrance, Alan Tracey, Richard Challis, Richard Durbin, Mark Blaxter

1Tree of Life, Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, CB10 1SA, UK
2Nottingham Trent University, School of Animal, Rural and Environmental Sciences, Nottingham, NG25 0QF, UK
3Red Squirrel Officer, The Wildlife Trust for Lancashire, Manchester and North Merseyside, The Barn, Berkeley Drive, Bamber Bridge, Preston, PR5 6BY, UK
4Department of Life Sciences, Natural History Museum, London, SW7 5BD, UK

First published: 13 Feb 2020, 5:27
https://doi.org/10.12688/wellcomeopenres.15721.1
Latest published: 13 Feb 2020, 5:27
https://doi.org/10.12688/wellcomeopenres.15721.1

Abstract
We present a genome assembly from an individual male *Sciurus carolinensis* (the eastern grey squirrel; Vertebrata; Mammalia; Eutheria; Rodentia; Sciuridae). The genome sequence is 2.82 gigabases in span. The majority of the assembly (92.3%) is scaffolded into 21 chromosomal-level scaffolds, with both X and Y sex chromosomes assembled.

Keywords
Sciurus carolinensis, grey squirrel, genome sequence, chromosomal

This article is included in the Tree of Life gateway.

Open Peer Review

Invited Reviewers
1
2
version 1
13 Feb 2020

1. Erik Garrison, University of California, Santa Cruz, Santa Cruz, USA
2. Takafumi Katsumura, Kitasato University School of Medicine, Sagamihara, Japan

Any reports and responses or comments on the article can be found at the end of the article.
**Species taxonomy**
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomia; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia; Sciuromorpha; Sciuridae; Sciurinae; Sciurini; Sciurus; *Sciurus carolinensis* Gmelin, 1788 (NCBI txid 30640).

**Background**
The eastern grey squirrel, *Sciurus carolinensis*, is native to eastern North America, where it plays important roles in forest regeneration through its habit of caching food nuts and seeds (Corbet & Hill, 1991). In North America, *S. carolinensis* has been introduced outside its native range such that it is now found from the Canadian Pacific northwest to Florida. *S. carolinensis* was introduced to Britain (in 1876), Ireland (in 1911), Italy (in 1948), South Africa (before 1900), Australia (in 1880s, extirpated in 1973) and Pitcairn island (in 1987) (see https://www.cabi.org/isc/datasheet/49075). *S. carolinensis*, which thrives in urban parklands and gardens, is classed as invasive in Europe and on Pitcairn island. In Britain and Ireland the expansion of *S. carolinensis* populations has driven decline in populations of the native red squirrel, *Sciurus vulgaris*, which we have also assembled (Mead et al., 2020). The negative impact of *S. carolinensis* is through interspecific competition, leading to competitive exclusion of *S. vulgaris*, and by their carriage of squirrelpox virus, to which they are resistant but not *S. vulgaris* are not (Chantrey et al., 2014) (Darby et al., 2014). The *S. carolinensis* genome will aid analyses of resistance and susceptibility to squirrelpox, as well as to the genomics of invasiveness.

**Genome sequence report**
The genome was sequenced from DNA extracted from a naturally deceased male *S. carolinensis* collected as part of a squirrel monitoring project run by the Wildlife Trust for Lancashire, Manchester and North Merseyside. A total of 74-fold coverage in Pacific Biosciences single-molecule long reads (N50 28 kb) and 40-fold coverage in 10X Genomics read clouds (from molecules with an estimated N50 of 19 kb) were generated. Primary assembly contigs were scaffolded with chromosome conformation HiC data (42-fold coverage). A contamination check identified a small number of low-coverage contigs that were likely to have derived from an apicomplexan parasite infecting the squirrel (Léveillé et al., 2020); these were removed. Subsequent manual assembly curation corrected 272 missing/missjoins and removed three haplotype duplications, reducing the scaffold number by 19% and increasing the scaffold N50 by 242% The final assembly has a total length of 2.82 Gb in 752 sequence scaffolds with a scaffold N50 of 148.2 Mb (Table 1). The majority, 92.3%, of the assembly sequence was assigned to some (numbered by sequence length), and the X and Y sex chromosomes (Figure 1–Figure 5; Table 2) plus 13 unlocalised scaffolds (assigned to chromosomes but with ambiguous placement). The assembly has a BUSCO (Simão et al., 2015) completeness of 93.7% 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 *S. carolinensis* mSciCar1 genome sequence is largely collinear with that of *S. vulgaris* mSciVul1 (Figure 4).

**Methods**
The eastern grey squirrel specimen was collected by the Wildlife Trust for Lancashire, Manchester and North Merseyside as part of an ongoing programme of recovery of dead squirrels. A full tissue dissection 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

---

1 See https://animaldiversity.org/accounts/Sciurus_carolinensis/
DNA Isolation Soft Tissue Protocol\(^2\). 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 I (single molecule long read) and Illumina HiSeq X (10X Genomics Chromium). HiC data were generated using the Dovetail v1.0 kit and sequenced on HiSeq X.

See Table 3 for software versions and sources. Assembly was carried out 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 was carried out using SALSA2. 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

Figure 2. Genome assembly of *Sciurus carolinensis* mSciCar1: GC-coverage plot. BlobToolKit GC-coverage plot of *S. carolinensis* mSciCar1 from long read data submission ERR3316154. The interactive version is available here.

Figure 3. Genome assembly of *Sciurus carolinensis* mSciCar1: Cumulative sequence plot. The blue line in the main plot shows the cumulative sequence plot for mSciCar. The sashed line shows the cumulative sequence plot of *S. vulgaris* mSciVu1 for comparison. The interactive version is available here.
**Figure 4.** Genome assembly of *Sciurus carolinensis* mSciCar1: Whole genome alignment with *Sciurus vulgaris* mSciVul1. A nucmer (Kurtz *et al.*, 2004) pairwise alignment of mSciCar1 (x-axis) with mSciVul1 (Y axis).

**Figure 5.** Genome assembly of *Sciurus carolinensis* mSciCar1: Hi-C contact map. Hi-C scaffolding of the *S. carolinensis* mSciCar1 assembly visualised in HiGlass (Kerpedjiev *et al.*, 2018).
Table 2. Chromosomal pseudomolecules in the genome assembly of Sciurus carolinensis mSciCar1.

<table>
<thead>
<tr>
<th>ENA accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR738590.1</td>
<td>1</td>
<td>208.99</td>
<td>40.3</td>
</tr>
<tr>
<td>LR738591.1</td>
<td>2</td>
<td>199.83</td>
<td>40.8</td>
</tr>
<tr>
<td>LR738592.1</td>
<td>3</td>
<td>183.55</td>
<td>40.3</td>
</tr>
<tr>
<td>LR738593.1</td>
<td>4</td>
<td>177.11</td>
<td>39.5</td>
</tr>
<tr>
<td>LR738594.1</td>
<td>5</td>
<td>175.91</td>
<td>39.1</td>
</tr>
<tr>
<td>LR738595.1</td>
<td>6</td>
<td>162.27</td>
<td>38.7</td>
</tr>
<tr>
<td>LR738596.1</td>
<td>7</td>
<td>154.99</td>
<td>39.1</td>
</tr>
<tr>
<td>LR738597.1</td>
<td>8</td>
<td>148.23</td>
<td>40.5</td>
</tr>
<tr>
<td>LR738598.1</td>
<td>9</td>
<td>141.42</td>
<td>38.8</td>
</tr>
<tr>
<td>LR738599.1</td>
<td>10</td>
<td>140.98</td>
<td>38.1</td>
</tr>
<tr>
<td>LR738600.1</td>
<td>11</td>
<td>135.23</td>
<td>40.1</td>
</tr>
<tr>
<td>LR738602.1</td>
<td>12</td>
<td>118.65</td>
<td>40.1</td>
</tr>
<tr>
<td>LR738603.1</td>
<td>13</td>
<td>94.68</td>
<td>41.1</td>
</tr>
<tr>
<td>LR738604.1</td>
<td>14</td>
<td>88.65</td>
<td>40.2</td>
</tr>
<tr>
<td>LR738605.1</td>
<td>15</td>
<td>83.14</td>
<td>40.5</td>
</tr>
<tr>
<td>LR738606.1</td>
<td>16</td>
<td>68.57</td>
<td>44.7</td>
</tr>
<tr>
<td>LR738607.1</td>
<td>17</td>
<td>66.05</td>
<td>42.7</td>
</tr>
<tr>
<td>LR738608.1</td>
<td>18</td>
<td>41.56</td>
<td>47.8</td>
</tr>
<tr>
<td>LR738609.1</td>
<td>19</td>
<td>30.99</td>
<td>44</td>
</tr>
<tr>
<td>LR738601.1</td>
<td>X</td>
<td>131.72</td>
<td>37.8</td>
</tr>
<tr>
<td>LR738610.1</td>
<td>Y</td>
<td>4.81</td>
<td>38.3</td>
</tr>
<tr>
<td></td>
<td>Unplaced</td>
<td>258.08</td>
<td>40</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>falcon-kit 1.2.2</td>
<td>(Chin et al., 2016)</td>
</tr>
<tr>
<td>purge_dups</td>
<td>1.0.0</td>
<td>(Guan et al., 2020)</td>
</tr>
<tr>
<td>SALSA2</td>
<td>2.2</td>
<td>(Ghurye et al., 2018)</td>
</tr>
<tr>
<td>scaffoldX</td>
<td>4.2</td>
<td><a href="https://github.com/wtsi-hpag/Scaffold10X">https://github.com/wtsi-hpag/Scaffold10X</a></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>lonrangeralign</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.html">http://samtools.github.io/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>
<tr>
<td>nucmer from MUMmer 3</td>
<td>3.0</td>
<td>(Kurtz et al., 2004)</td>
</tr>
</tbody>
</table>

The genome sequence is released openly for reuse. The S. carolinensis genome sequencing initiative is part of the Wellcome Sanger Institute’s “25 genomes for 25 years” project. It is also part of the Vertebrate Genome Project (VGP) and the Darwin Tree of Life (DToL) project. The specimen has been preserved in ethanol and deposited with the Natural History Museum, London under registration number NHMUK ZD 2019.214, where it will remain accessible to the research community for posterity. All raw 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.

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.

3 https://www.sanger.ac.uk/science/collaboration/25-genomes-25-years
4 https://vertebrategenomesproject.org/
5 https://www.darwintreeoflife.org/
Pubmed Abstract | Publisher Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
Publisher Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
Reference Source
Publisher Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text | Free Full Text
PubMed Abstract | Publisher Full Text
Publisher Full Text
PubMed Abstract | Publisher Full Text
Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 09 November 2020

https://doi.org/10.21956/wellcomeopenres.17234.r40853

© 2020 Katsumura T. 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.

Takafumi Katsumura
Department of Anatomy, Kitasato University School of Medicine, Sagamihara, Japan

The authors constructed a whole-genome sequence for the eastern gray squirrel using spleen-derived DNA from naturally dead squirrels collected in a Wildlife Trust project. This genome sequence is based on long reads (74-fold coverage) from the PacBio sequencer, short reads from 10X Genomics + Illumina sequencer (40-fold coverage), and scaffolding by Hi-C data (42-fold coverage). From the various genome statistics calculated and BUSCO's score, I think this whole-genome sequence is useful for comparative genome studies.

My comments are as follows:

1. I would suggest that the authors describe the age of the squirrel used as samples (if the authors know) and the collecting date. This information may be useful for future secondary use of the Hi-C data acquired in this paper. Also, I would suggest that the authors describe how the amount was the spleen used for DNA extraction.

2. I would suggest that the authors add the number of libraries used to generate the reads by the PacBio and Illumina in the text, respectively. The readers could know their numbers by looking and counting them in Table 1, but I think it will help the reader understand this work's data quality if specified in the text.

3. For the contact map in Figure 5, the authors should describe the X and Y axes and label them. Also, I think it would be easier for the reader to understand the figure if there is a color key.

4. To drive the reproducibility of the data by readers, it would be important to describe the settings of the software tools used to construct the genome sequence. I would suggest that the authors describe the settings of all software tools in the text.

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:** Evolutionary 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 17 February 2020

https://doi.org/10.21956/wellcomeopenres.17234.r37915

© 2020 Garrison E. 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.

✅ **Erik Garrison**
UC Santa Cruz Genomics Institute, University of California, Santa Cruz, Santa Cruz, CA, USA

The authors present the assembly of a male eastern grey squirrel, based on PacBio single molecule sequencing and 10X genomics linked read sequencing. The approach is technically very sound, and mirrors that used in the Vertebrate Genomes Project. The assembly’s completeness is impressive, as the authors show by comparison to an existing assembly from the red squirrel, evaluation of the assembly against BUSCO, and several plots from BlobToolKit. I am confident that this assembly will be usable by researchers working on this species.

I would suggest that the authors improve the rendering of several of the figures. Those produced by the BlobToolKit have very small font relative to their rasterized pixel density. I would either render them as vector graphics or adjust the rendering (if possible) to improve the font size. The HiGlass plot clearly demonstrates the expected pattern of connectivity across the chromosome-scale scaffolds, but overplotting of the delimiting line (grey bars) makes the region of the plot (to the bottom and right) referring to the smaller scaffolds completely illegible. If this can be fixed, it might make the plot a little more interesting.

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 usable and accessible format? Yes

Competing Interests: I have published work with some of the authors before but I do not believe this affected my ability to review impartially.

Reviewer Expertise: (pan)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.