DATA NOTE

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

Dan Mead[1], Kathryn Fingland[2], Rachel Cripps[3], Roberto Portela Miguez[4], Michelle Smith[1], Craig Corton[1], Karen Oliver[1], Jason Skelton[1], Emma Betteridge[1], Jale Doulcan[1], Michael A. Quail[1], Shane A. McCarthy[1], Kerstin Howe[1], Ying Sims[1], James Torrance[1], Alan Tracey[1], Richard Challis[1], Richard Durbin[1], Mark Blaxter[1]

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

v1 **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

Open Peer Review

Reviewer Status ✓ ✓

	Invited Reviewers	
	1	**2**
version 1 13 Feb 2020	✓ report	✓ report

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.

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

Author roles: Mead D: Conceptualization, Investigation, Writing – Review & Editing; **Fingland K**: Formal Analysis, Investigation, Resources, Writing – Review & Editing; **Cripps R**: Formal Analysis, Investigation, Resources, Writing – Review & Editing; **Portela Miguez R** : 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, Supervision, Writing – Review & Editing; **Skelton J**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Betteridge E**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Doulcan J**: Writing – Review & Editing; **Quail MA**: Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **McCarthy SA**: Data Curation, Formal Analysis, Investigation, Software, Writing – Review & Editing; **Howe K**: Data Curation, Formal Analysis, Investigation, Resources, Software, Supervision, Writing – Review & Editing; **Sims Y**: Data Curation, Formal Analysis, Investigation, Software, Validation, Writing – Review & Editing; **Torrance J**: Data Curation, Formal Analysis, Investigation, Validation, Writing – Review & Editing; **Tracey A**: Data Curation, Formal Analysis, Investigation, Software, Writing – Review & Editing; **Challis R**: Data Curation, Formal Analysis, Investigation, Software, Validation, Visualization, Writing – Review & Editing; **Durbin R**: Conceptualization, Investigation, Resources, Supervision, Writing – Review & Editing; **Blaxter M**: Conceptualization, Funding Acquisition, Project Administration, Supervision, 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 (WT206194). SMcC and RD were supported by Wellcome grant WT207492. MB was supported through Wellcome grant WT218328.
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: Mead D, Fingland K, Cripps R *et al.* **The genome sequence of the eastern grey squirrel, *Sciurus carolinensis* Gmelin, 1788 [version 1; peer review: 2 approved]** Wellcome Open Research 2020, **5**:27
https://doi.org/10.12688/wellcomeopenres.15721.1

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

Species taxonomy

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; 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)[1]. 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 *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/misjoins and removed three haplotypic 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 21 chromosomal-level scaffolds representing 19 autosomes (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

Table 1. Genome data for Sciurus carolinensis mSciCar1.

Project accession data	
Assembly identifier	mSciCar1
Species	*Sciurus carolinensis*
Specimen	NHMUK ZD 2019.214
NCBI taxonomy ID	30640
BioProject	PRJEB35386
Biosample ID	SAMEA994726
Isolate information	Wild isolate; male
Raw data accessions	
PacificBiosciences SEQUEL I	ERR3313242-ERR3313245, ERR3313247-ERR3313255, ERR3313329, ERR3313331, ERR3313332, ERR3313342-ERR3313348
10X Genomics Illumina	ERR3316153-ERR3316156, ERR3316173-ERR3316176
Hi-C Illumina	ERR3312499-ERR3312500, ERR3850937
Genome assembly	
Assembly accession	GCA_902686445.1
Accession of alternate haplotype	GCA_902685475.1
Span (Mb)	2,815,397,268
Number of contigs	2576
Contig N50 length (Mb)	13.98
Number of scaffolds	752
Scaffold N50 length (Mb)	148.23
Longest scaffold (Mb)	208.99
BUSCO* genome score	C:93.7%[S:92.3%,D:1.4%],F:2.8%, M:3.5%,n:4104

* BUSCO scores based on the mammalia_odb9 BUSCO set using v3.0.2. 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/mSciCar1_1/dataset/mSciCar1_1/busco

[1] See https://animaldiversity.org/accounts/Sciurus_carolinensis/

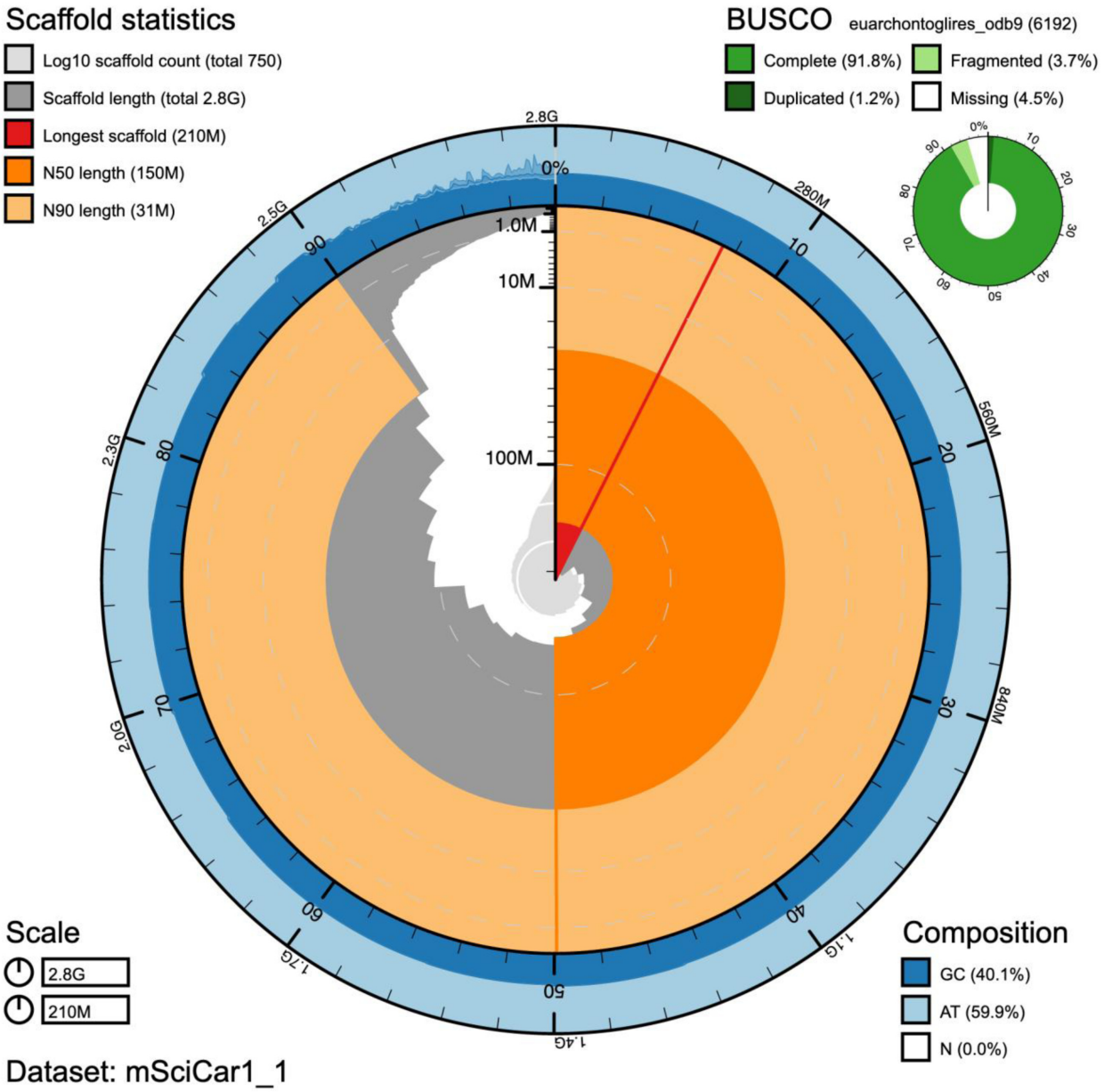

Figure 1. Genome assembly of *Sciurus carolinensis* mSciCar1: Metrics. BlobToolKit Snailplot showing N50 metrics for *S. carolinensis* assembly mSciCar1 and BUSCO scores for the Euarchontoglires set of orthologues. The interactive version is available here.

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

[2] https://bionanogenomics.com/wp-content/uploads/2018/02/30077-Bionano-Prep-Animal-Tissue-DNA-Isolation-Soft-Tissue-Protocol.pdf

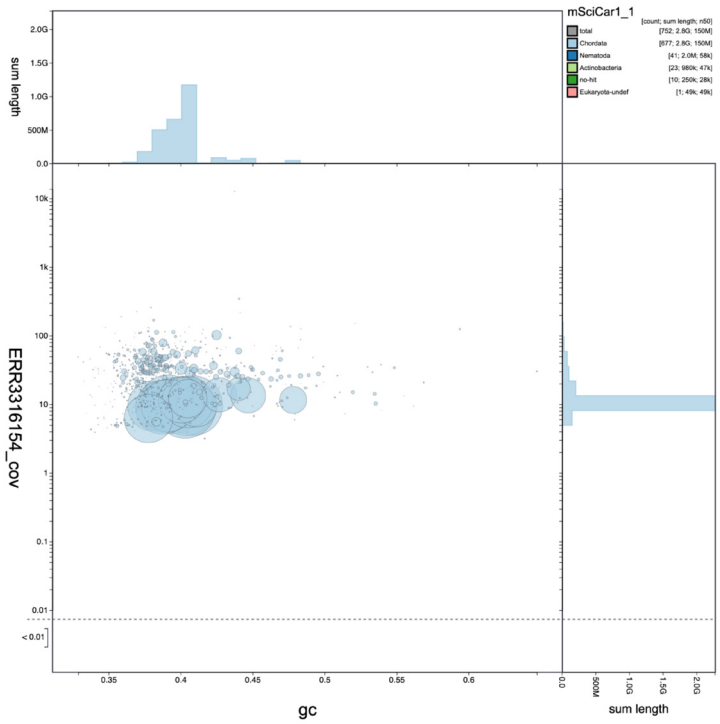

Figure 2. Genome assembly of *Sciurus carolinensis* mSciCar1: GC-coverage plot. BlobToolKit GC-coverage plot of *S. carolinensis* mSciCar1 from long read data submission ERR3316154. The interactive version is available here.

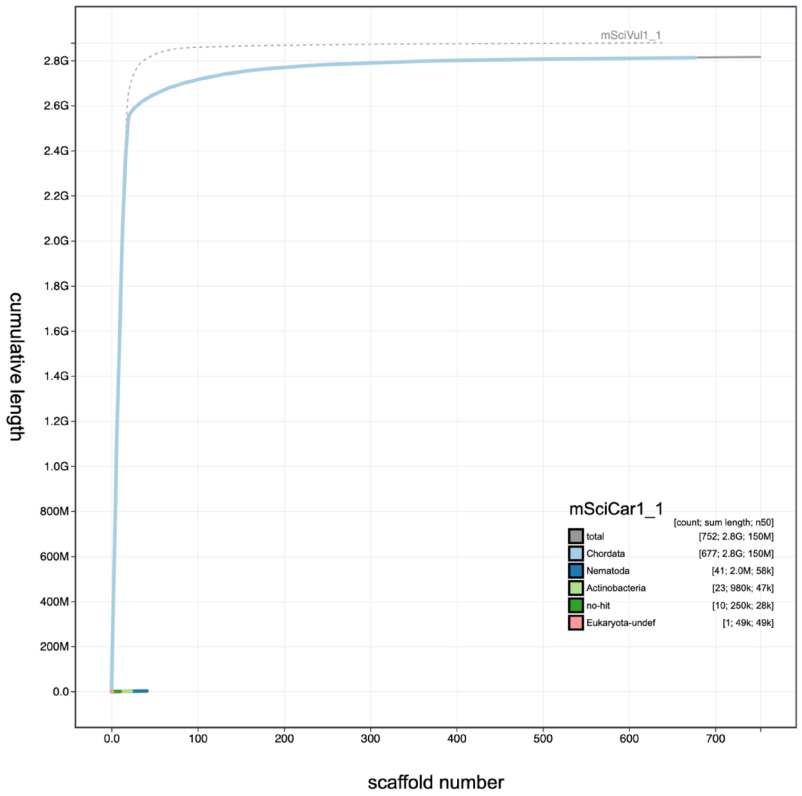

Figure 3. Genome assembly of *Sciurus carolinensis* mSciCar1: Cumulative sequence plot. The blue line in the main plot shows the cumulative sequence plot for mSciCar. The sashed line shows the cumulative sequence plot of *S. vulgaris* mSciVul1 for comparison. The interactive version is available here.

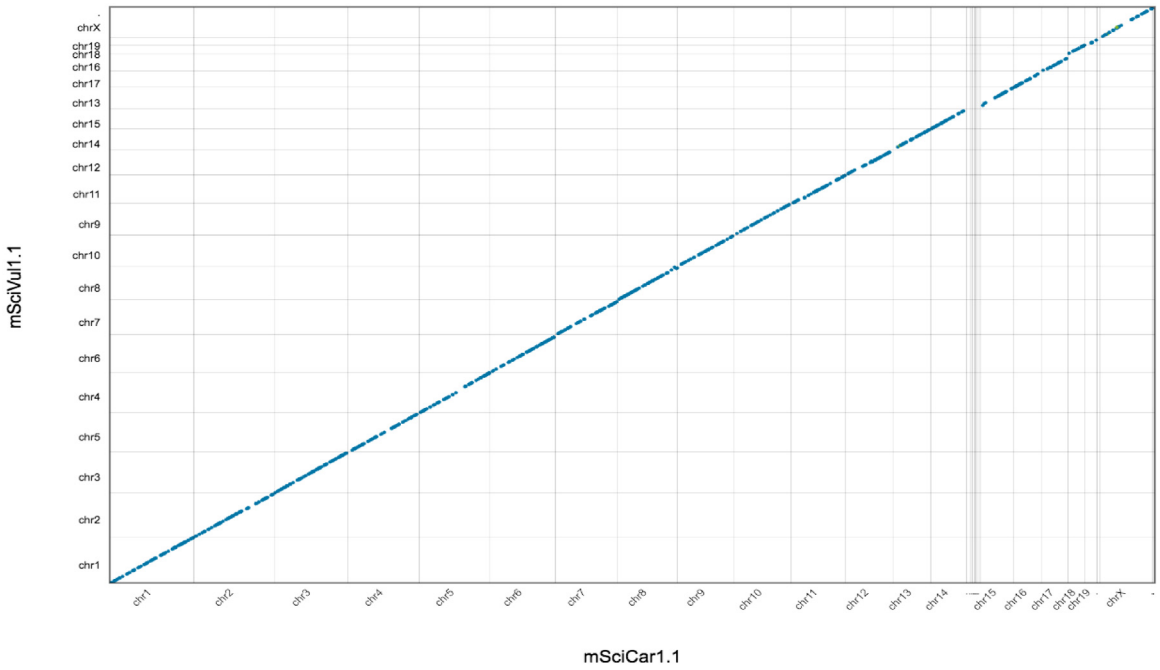

Figure 4. Genome assembly of *Sciurus carolinensis* mSciCar1: Whole genome alignment with *Sciurus vulgaris* mSciVul1. A nucmer (Kurtz *et al.*, 2004) pairwise alignment of mSciCar1 (x-axis) with mSciVul1 (Y axis).

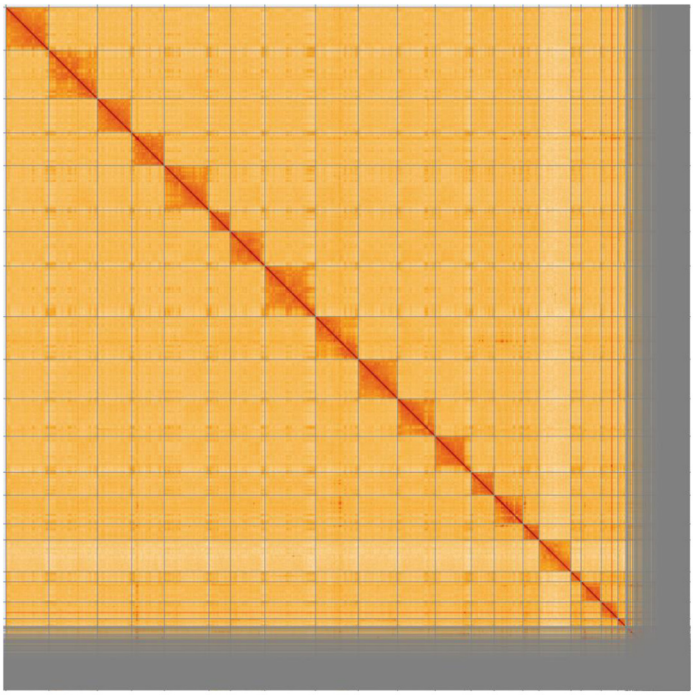

Figure 5. Genome assembly of *Sciurus carolinensis* mSciCar1: Hi-C contact map. Hi-C scaffolding of the *S. carolinensis* mSciCar1 assembly visualised in HiGlass (Kerpedjiev *et al.*, 2018).

Table 2. Chromosomal pseudomolecules in the genome assembly of *Sciurus carolinensis* mSciCar1.

ENA accession	Chromosome	Size (Mb)	GC%
LR738590.1	1	208.99	40.3
LR738591.1	2	199.83	40.8
LR738592.1	3	183.55	40.3
LR738593.1	4	177.11	39.5
LR738594.1	5	175.91	39.1
LR738595.1	6	162.27	38.7
LR738596.1	7	154.99	39.1
LR738597.1	8	148.23	40.5
LR738598.1	9	141.42	38.8
LR738599.1	10	140.98	38.1
LR738600.1	11	135.23	40.1
LR738602.1	12	118.65	40.1
LR738603.1	13	94.68	41.1
LR738604.1	14	88.65	40.2
LR738605.1	15	83.14	40.5
LR738606.1	16	68.57	44.7
LR738607.1	17	66.05	42.7
LR738608.1	18	41.56	47.8
LR738609.1	19	30.99	44
LR738601.1	X	131.72	37.8
LR738610.1	Y	4.81	38.3
-	Unplaced	258.08	40

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). Since Hi-C data were sparse, curation was aided by synteny with the assembly for *Sciurus vulgaris* simultaneously being curated by the Wellcome Sanger Institute. The genome was analysed within the BlobToolKit environment (Challis *et al.*, 2019).

Data availability

Underlying data

European Nucleotide Archive: *Sciurus carolinensis* (grey squirrel) genome assembly, mSciCar1. BioProject accession number PRJEB35386; https://identifiers.org/ena.embl:PRJEB35386.

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)
SALSA2	2.2	(Ghurye *et al.*, 2018)
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)
BlobToolKit	1	(Challis *et al.*, 2019)
nucmer from MUMmer 3	3.0	(Kurtz *et al.*, 2004)

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[3]. It is also part of the Vertebrate Genome Project (VGP)[4] and the Darwin Tree of Life (DToL) project[5]. 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/

References

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – Interactive Quality Assessment of Genome Assemblies.** *bioRxiv.* 2019.
Publisher Full Text

Chantrey J, Dale TD, Read JM, *et al.*: **European red squirrel population dynamics driven by squirrelpox at a gray squirrel invasion interface.** *Ecol Evol.* 2014; **4**(19): 3788–99.
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

Corbet GB, Hill JE: **A World List of Mammalian Species (3rd Edition).** Natural History Museum Publications/Oxford University Press, 1991, 243 Pp., HB £30.00. *Oryx.* Cambridge University Press, 1991; **25**(3): 174.
Publisher Full Text

Darby AC, McInnes CJ, Kjær KH, *et al.*: **Novel host-related virulence factors are encoded by squirrelpox virus, the main causative agent of epidemic disease in red squirrels in the UK.** *PLoS One.* 2014; **9**(7): e96439.
PubMed Abstract | Publisher Full Text | Free Full Text

Garrison E, Marth G: **Haplotype-Based Variant Detection from Short-Read Sequencing.** *arXiv [q-bio.GN].* arXiv, 2012.
Reference Source

Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C Links with Assembly Graphs for Chromosome-Scale Assembly.** *bioRxiv.* 2018.
Publisher Full Text

Ghurye J, Rhie A, Walenz BP, *et al.*: **Integrating Hi-C links with assembly graphs for chromosome-scale assembly.** *PLoS Comput Biol.* 2019; **15**(8): e1007273.
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; pii: btaa025.
PubMed Abstract | 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

Kurtz S, Phillippy A, Delcher AL, *et al.*: **Versatile and open software for comparing large genomes.** *Genome Biol.* 2004; **5**(2): R12.
PubMed Abstract | Publisher Full Text | Free Full Text

Léveillé, AN, El Skhawy N, Barta JR: **Multilocus sequencing of *Hepatozoon* cf. *griseisciuri* infections in Ontario eastern gray squirrels (*Sciurus carolinensis*) uncovers two genotypically distinct sympatric parasite species.** *Parasitol Res.* 2020; **119**(2): 713–724.
PubMed Abstract | Publisher Full Text

Mead D, Fingland K, Cripps R, *et al.*: **The genome sequence of the Eurasian red squirrel, *Sciurus vulgaris* Linnaeus 1758 [version 1; peer review: awaiting peer review].** *Wellcome Open Res.* 2020; **5**: 18.
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

Open Peer Review

Current Peer Review Status: ✔ ✔

Version 1

Reviewer Report 09 November 2020

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

✔ **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

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 useable 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.