The genome sequence of the common pipistrelle, *Pipistrellus pipistrellus* Schreber 1774 [version 1; peer review: 2 approved with reservations]

Christopher Vine\textsuperscript{1}, Emma C. Teeling\textsuperscript{2}, Michelle Smith\textsuperscript{3}, Craig Corton\textsuperscript{3}, Karen Oliver\textsuperscript{3}, Jason Skelton\textsuperscript{3}, Emma Betteridge\textsuperscript{3}, Jale Doulcan\textsuperscript{3,4}, Michael A. Quail\textsuperscript{3}, Shane A. McCarthy\textsuperscript{3,5}, Kerstin Howe\textsuperscript{3}, James Torrance\textsuperscript{3}, Jonathan Wood\textsuperscript{3}, Sarah Pelan\textsuperscript{3}, Ying Sims\textsuperscript{3}, Richard Challis\textsuperscript{3}, Jonathan Threlfall\textsuperscript{3}, Daniel Mead\textsuperscript{3,6}, Mark Blaxter\textsuperscript{3}

\textsuperscript{1}Cambridgeshire Bat Group, Cambridgeshire, UK  
\textsuperscript{2}School of Biology and Environmental Science, University College Dublin, Belfield, Dublin 4, Ireland  
\textsuperscript{3}Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK  
\textsuperscript{4}Achilles Therapeutics plc, London, W6 8PW, UK  
\textsuperscript{5}Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK  
\textsuperscript{6}Owlstone Medical, Cambridge Science Park, Cambridge, CB4 0GJ, UK

**Abstract**
We present a genome assembly from an individual female *Pipistrellus pipistrellus* (the common pipistrelle; Chordata; Mammalia; Chiroptera; Vespertilionidae). The genome sequence is 1.76 gigabases in span. The majority of the assembly is scaffolded into 21 chromosomal pseudomolecules, with the X sex chromosome assembled.

**Keywords**
Pipistrellus pipistrellus, common pipistrelle, genome sequence, chromosomal

This article is included in the Tree of Life gateway.
Species taxonomy
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomia; Mammalia; Eutheria; Laurasiatheria; Chiroptera; Microchiroptera; Vespertilionidae; Pipistrellus; Pipistrellus pipistrellus Schreber 1774 (NCBI:txid59474).

Introduction
The common pipistrelle, Pipistrellus pipistrellus, is a small species of bat with a range that extends across Europe and into Central Asia and North Africa (Boston et al., 2014). It is one of the most common species of bat in the United Kingdom and Ireland, where it is considered to be of least concern on the Mammal Society’s Red List of extinction risk (Mathews & Harrower, 2020). Despite a decline in roost count, the population of common pipistrelles has increased since 1999 (Bat Conservation Trust, 2020). P. pipistrellus roosts in trees and buildings, emerging at dusk to feed on small flying insects, using laryngeal echolocation to orient and locate prey.

Originally thought to be a single species, it was not until fairly recently that the cryptic species P. pipistrellus and Pipistrellus pygmaeus (the soprano pipistrelle) were confirmed to be distinct (Barlow & Jones, 1999), although differences between the two species had been noted previously (Jones & van Parijs, 1993). The two species appear morphologically similar, but exhibit differences in their echolocation call peak frequency: P. pipistrellus (~45 kHz) and P. pygmaeus (~55 kHz). They exhibit small but significant genetic differences although hybridization between the two species has been observed (Sztencel-Jabłonka & Bogdanowicz, 2012).

This genome sequence will be of utility to researchers that wish to examine in depth the genetic differences between P. pipistrellus and its cryptic partner P. pygmaeus that underlie the small but significant divergence between the species.

Genome sequence report
The genome was sequenced from a single female P. pipistrellus collected from Potten End, Berkhamsted, Hertfordshire, UK. A total of 56-fold coverage in Pacific Biosciences single-molecule long reads (N50 19 kb) and 34-fold coverage in 10X Genomics read clouds (from molecules with an estimated N50 of 31 kb) were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data. Manual assembly curation corrected 408 missing/misjoins and removed 19 haplotypic duplications, reducing the scaffold number by 35.5%, increasing the scaffold N50 by 126.2% and decreasing the assembly length by 0.1%. The final assembly has a total length of 2.88 Gb in 323 sequence scaffolds with a scaffold N50 of 94.9 Mb (Table 1). The majority, 98.8%, of the assembly sequence was assigned to 22 chromosomal-level scaffolds representing 21 autosomes (numbered by sequence length), and the X sex chromosome (Figure 1–Figure 4; Table 2). The assembly has a BUSCO (Simão et al., 2015) completeness of 89.6% using the mammalia_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 Pipistrellus pipistrellus, mPipPip1.1.

<table>
<thead>
<tr>
<th>Project accession data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly identifier</td>
<td>mPipPip1.1</td>
</tr>
<tr>
<td>Species</td>
<td>Pipistrellus pipistrellus</td>
</tr>
<tr>
<td>Specimen</td>
<td>mPipPip1 (Genome assembly); mPipPip2 (Hi-C)</td>
</tr>
<tr>
<td>NCBI taxonomy ID</td>
<td>NCBI:txid59474</td>
</tr>
<tr>
<td>BioProject</td>
<td>PRJEB39564</td>
</tr>
<tr>
<td>BioSample ID</td>
<td>SAMEA994724</td>
</tr>
<tr>
<td>Isolate information</td>
<td>Female, spleen tissue</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw data accessions</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>PacificBiosciences SEQUEL I</td>
<td>ERX3338750-ERX3338753, ERX3338772, ERX3338784, ERX3338787, ERX3338792, ERX3338841, ERX3338842, ERX3338876, ERX3338877</td>
</tr>
<tr>
<td>10X Genomics Illumina</td>
<td>ERX3341659-ERX3341662, ERX3341687-ERX3341690</td>
</tr>
<tr>
<td>Hi-C Illumina</td>
<td>ERX5308917, ERX5308918</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome assembly</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly accession</td>
<td>GCA_903992545.1</td>
</tr>
<tr>
<td>Accession of alternate haplotype</td>
<td>GCA_903992515.1</td>
</tr>
<tr>
<td>Span (Mb)</td>
<td>1,763</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>1,868</td>
</tr>
<tr>
<td>Contig N50 length (Mb)</td>
<td>4446.8</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>323</td>
</tr>
<tr>
<td>Scaffold N50 length (Mb)</td>
<td>94.9</td>
</tr>
<tr>
<td>Longest scaffold (Mb)</td>
<td>203.6</td>
</tr>
</tbody>
</table>

* BUSCO scores based on the mammalia_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 genomedhubs.org/view/Pipistrellus\%20pipistrellus/dataset/CAJEUD01/busc0.
**Methods**

The common pipistrelle specimen was a female individual found at a residential address in Potten End, Berkhamsted, Hertfordshire, UK. The animal had died during renovation of a private home during works licensed by Natural England under the Bat Low Impact Class License, WML-CL21.

DNA was extracted using an agarose plug extraction from spleen tissue 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 the manufacturers’ instructions. Hi-C data were generated using the Arima Hi-C kit from a separate
tissue sample (mPipPip2) taken from the same animal. Sequencing was performed by the Scientific Operations DNA Pipelines at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL I (long read) and Illumina HiSeq X (10X, Hi-C) instruments.

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 removed with purge_dups (Guan et al., 2020) and a first round of scaffolding carried out with 10X Genomics read clouds using scaff10x (see Table 3 for software versions and sources). 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

Figure 2. Genome assembly of *Pipistrellus pipistrellus*, mPipPip1.1: GC coverage. BlobToolKit GC-coverage plot. The “arthropod” matches are segments that have marginal best hits to Apidae proteins and are unlikely to be contamination. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Pipistrellus%20pipistrellus/dataset/CAJEUD01/blob?plotShape=circle.
Figure 3. Genome assembly of *Pipistrellus pipistrellus*, mPipPip1.1: cumulative sequence. BlobToolKit cumulative sequence plot. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/Pipistrellus%20pipistrellus/dataset/CAJEUD01/cumulative.

Figure 4. Genome assembly of *Pipistrellus pipistrellus*, mPipPip1.1: Hi-C contact map. Hi-C contact map of the mPipPip1.1 assembly, visualised in HiGlass.
Table 2. Chromosomal pseudomolecules in the genome assembly of *Pipistrellus pipistrellus*, mPipPip1.1.

<table>
<thead>
<tr>
<th>ENA accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR862356.1</td>
<td>1</td>
<td>203.62</td>
<td>40.9</td>
</tr>
<tr>
<td>LR862357.1</td>
<td>2</td>
<td>194.58</td>
<td>41.6</td>
</tr>
<tr>
<td>LR862358.1</td>
<td>3</td>
<td>188.94</td>
<td>40.6</td>
</tr>
<tr>
<td>LR862360.1</td>
<td>4</td>
<td>100.18</td>
<td>41.9</td>
</tr>
<tr>
<td>LR862361.1</td>
<td>5</td>
<td>94.93</td>
<td>43.6</td>
</tr>
<tr>
<td>LR862362.1</td>
<td>6</td>
<td>85.79</td>
<td>40.8</td>
</tr>
<tr>
<td>LR862363.1</td>
<td>7</td>
<td>84.55</td>
<td>43.5</td>
</tr>
<tr>
<td>LR862364.1</td>
<td>8</td>
<td>83.52</td>
<td>41.1</td>
</tr>
<tr>
<td>LR862365.1</td>
<td>9</td>
<td>72.54</td>
<td>43</td>
</tr>
<tr>
<td>LR862366.1</td>
<td>10</td>
<td>71.5</td>
<td>41.3</td>
</tr>
<tr>
<td>LR862367.1</td>
<td>11</td>
<td>69.39</td>
<td>42.8</td>
</tr>
<tr>
<td>LR862368.1</td>
<td>12</td>
<td>58.29</td>
<td>43.1</td>
</tr>
<tr>
<td>LR862369.1</td>
<td>13</td>
<td>53.28</td>
<td>44.2</td>
</tr>
<tr>
<td>LR862370.1</td>
<td>14</td>
<td>51.67</td>
<td>43.8</td>
</tr>
<tr>
<td>LR862371.1</td>
<td>15</td>
<td>47.28</td>
<td>42</td>
</tr>
<tr>
<td>LR862372.1</td>
<td>16</td>
<td>46.23</td>
<td>47.5</td>
</tr>
<tr>
<td>LR862373.1</td>
<td>17</td>
<td>45.57</td>
<td>47.5</td>
</tr>
<tr>
<td>LR862374.1</td>
<td>18</td>
<td>39.45</td>
<td>46.8</td>
</tr>
<tr>
<td>LR862375.1</td>
<td>19</td>
<td>24.19</td>
<td>49.9</td>
</tr>
<tr>
<td>LR862376.1</td>
<td>20</td>
<td>15.57</td>
<td>46.5</td>
</tr>
<tr>
<td>LR862377.1</td>
<td>21</td>
<td>7.78</td>
<td>51.2</td>
</tr>
<tr>
<td>LR862359.1</td>
<td>X</td>
<td>102.77</td>
<td>39.9</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 1.8.0</td>
<td>(Chin et al., 2016)</td>
</tr>
<tr>
<td>purge_dups</td>
<td>1.2.3-b542dbf</td>
<td>(Guan et al., 2020)</td>
</tr>
<tr>
<td>SALSA2</td>
<td>2.2-14-g974589f</td>
<td>(Ghurye et al., 2019)</td>
</tr>
<tr>
<td>scaffold10x</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>arrow</td>
<td>gcpps 1.9.0-SL-release-8.0.0+1-37-gd7b188d</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>1.3.1-17-gaa2ace8</td>
<td>(Garrison &amp; Marth, 2012)</td>
</tr>
<tr>
<td>bcftools consensus</td>
<td>1.9-78-gb7e4ba9</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></td>
<td>(Chow et al., 2016)</td>
</tr>
<tr>
<td>HiGlass</td>
<td>1.11.6</td>
<td>(Kerpedjiev et al., 2018)</td>
</tr>
<tr>
<td>PretextView</td>
<td>0.0.4</td>
<td><a href="https://github.com/wtsi-hpag/PretextView">https://github.com/wtsi-hpag/PretextView</a></td>
</tr>
<tr>
<td>BlobToolKit</td>
<td>1.2</td>
<td>(Challis et al., 2020)</td>
</tr>
</tbody>
</table>
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. Manual curation was performed as described previously (Howe et al., 2021) using the gEVAL system (Chow et al., 2016), Bionano Access, HiGlass and Pre-text. Figure 1–Figure 3 were generated using BlobToolKit (Challis et al., 2020).

Data availability
Underlying data
European Nucleotide Archive: mPipPip1 (common pipistrelle), Accession number PRJEB39564: https://www.ebi.ac.uk/ena/browser/view/PRJEB39566

The genome sequence is released openly for reuse. The *P. pipistrellus* 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, the Darwin Tree of Life (DTOL) project, and Bat1K. 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.

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

References


Open Peer Review

Current Peer Review Status: ?  ?

Version 1

Reviewer Report 14 June 2021

https://doi.org/10.21956/wellcomeopenres.18638.r43926

© 2021 Panitz F. 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.

Frank Panitz
Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

The genome sequence of the bat pipistrelle will provide a valuable tool to investigate the genetics between bats and related species. The authors apply current state-of-the-art methods to generate an assembly resulting in chromosome-level pseudomolecules. Still, some aspects regarding the assembly quality have to be re-assessed and improved.

- An account of the estimated (haploid) genome size and how it relates to the assembled sequences is missing. K-mer analysis to identify k-mers that are overrepresented as well as an evaluation of repeat sequences would help to better assess the resulting assembly.

- The contig N50 length given in table 1 is larger than the scaffold N50 values, which does not appear normal.

- The coverage of the autosome is generally between 14 and 15-fold; for the chromosomal pseudomolecule 19, however, coverage is more than twice as high (> 33-fold; https://blobtoolkit.genomehubs.org/view/Pipistrellus%20pipistrellus/dataset/CAJEUD01/table?ERR3316149_cov-Active=true&ERR3316178_cov--Active=true#Lists). The authors should explain this observation; as high levels of repeats, that fail to assemble correctly may account for this discrepancy an assessment or annotation of repetitive elements should be provided. Backmapping of short (Illumina) reads is suggested to corroborate the coverage distribution over the genome.

- The BUSCO assessment presented indicates that about 10% of the BUSCO genes are not accounted for. The authors should comment on this aspect and might consider experimental validation e.g. by transcriptome sequencing.

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

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

**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:** Molecular 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, however I have significant reservations, as outlined above.

Reviewer Report 26 May 2021

https://doi.org/10.21956/wellcomeopenres.18638.r43923

© 2021 Wang W. 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.

Wen Wang
CAS-Max Planck Junior Research Group, State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan, 650223, China

The genome of *pipistrelle* will be a very important genetic resource for studies related to bat evolution. However, more procedures have to be done to assess the genome and to improve the data reliability.

1. The genome is lack of pre-assembly survey, as no k-mer analysis nor FCM was described. This makes it difficult to assess the genome assembly.

2. The methods to assess the genome assembly are too simple and not comprehensive enough. More methods such as the reads mapping ratio and QV score are needed. The information about gene prediction and annotation could not be found across the manuscript, which is important for further analysis. The quality of this genome needs to be assessed more comprehensively. For example, the Illumina reads need to be mapped back to the genome, and the mapping rate of the reads and whether they are evenly distributed across the genome need to be checked.

3. Some methods and figures are lack of statement and explanation. For example, the version of BUSCO was not mentioned. Figure 2 and Figure 3, these two figures offered me little information. What is ERR3316150_cov in figure2, and what is the purpose to draw it? Is it necessary to put 2 meaningless figs here to just show the genome GC content and
contamination respectively?

4. The completeness accessed by BUSCO is not that high. About 9.8% of the BUSCO genes are missing in the assembly. The author should provide an explain to this phenomenon and compare the genome quality with other bats. Whether this is a characteristic of the bat itself or the assembly needs further improvement.

5. In Table 1, the Contig N50 length is 4446.8 Mb. This is an abnormal number.

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

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

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 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, however I have significant reservations, as outlined above.