DATA NOTE

The genome sequence of the malaria mosquito, Anopheles funestus, Giles, 1900 [version 1; peer review: awaiting peer review]

Diego Ayala¹,², Ousman Akone-Ella², Pierre Kengne¹,², Harriet Johnson³, Haynes Heaton⁴, Joanna Collins⁵, Ksenia Krasheninnikova⁵, Sarah Pelan⁵, Damon-Lee Pointon⁵, Ying Sims⁵, James Torrance¹,², Alan Tracey⁵, Marcela Uliano-Silva⁵, Katharina von Wyschetzki⁵, Jonathan Wood¹,², DNA Pipelines collective, Shane McCarthy⁵,⁶, Daniel Neafsey⁷,⁸, Alex Makunin⁵, Mara Lawniczak¹⁵

¹MIVEGEC, IRD, Montpellier, 34394, France
²ESV-GAB, Centre Interdisciplinaire de Recherches Médicales de Franceville (CIRMF), Franceville, BP 769, Gabon
³Scientific Operations, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
⁴CSSE, Auburn University, Auburn, AL, 36849, USA
⁵Tree of Life, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK
⁶Department of Genetics, University of Cambridge, Cambridge, CB2 3EH, UK
⁷Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA
⁸Infectious Disease and Microbiome Program, Broad Institute, Cambridge, MA, 02142, USA

V1 First published: 25 Nov 2022, 7:287
https://doi.org/10.12688/wellcomeopenres.18445.1
Latest published: 25 Nov 2022, 7:287
https://doi.org/10.12688/wellcomeopenres.18445.1

Abstract
We present a genome assembly from an individual female Anopheles funestus (the malaria mosquito; Arthropoda; Insecta; Diptera; Culicidae). The genome sequence is 251 megabases in span. The majority of the assembly is scaffolded into three chromosomal pseudomolecules with the X sex chromosome assembled. The complete mitochondrial genome was also assembled and is 15.4 kilobases in length.

Keywords
Anopheles funestus, African malaria mosquito, genome sequence, chromosomal inversions

This article is included in the Tree of Life gateway.
Corresponding author: Mara Lawniczak (mara@sanger.ac.uk)

Author roles: Ayala D: Methodology, Writing – Review & Editing; Akone-Ella O: Methodology; Kengne P: Methodology; Johnson H: Methodology; Heaton H: Investigation, Methodology; Collins J: Data Curation; Krasheninnikova K: Methodology; Pelan S: Data Curation; Pointon DL: Data Curation; Sims Y: Data Curation; Torrance J: Data Curation; Tracey A: Data Curation; Uliano-Silva M: Methodology; von Wyschetzki K: Methodology; Wood J: Data Curation; McCarthy S: Methodology; Neafsey D: Conceptualization, Funding Acquisition; Makunin A: Formal Analysis, Methodology, Project Administration, Visualization, Writing – Original Draft Preparation; Lawniczak M: Conceptualization, Funding Acquisition, Investigation, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by a Bill & Melinda Gates Foundation Award (INV-009760) to MKNL. The Wellcome Sanger Institute is funded by Wellcome (206194, https://doi.org/10.35802/206194), which supports MKNL. Field collection was supported by an ANR grant (ANR-18-CE35-0002-01 –WILDING). SAM is supported from a Wellcome Grant (207492, https://doi.org/10.35802/207492). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Ayala 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: Ayala D, Akone-Ella O, Kengne P et al. The genome sequence of the malaria mosquito, Anopheles funestus, Giles, 1900 [version 1; peer review: awaiting peer review] Wellcome Open Research 2022, 7:287 https://doi.org/10.12688/wellcomeopenres.18445.1

First published: 25 Nov 2022, 7:287 https://doi.org/10.12688/wellcomeopenres.18445.1
Species taxonomy
Animalia; Arthropoda; Insecta; Diptera; Culicidae; Anophelinae; Anopheles; *Anopheles funestus*; Giles, 1900 (NCBI txid:62324).

Background
The mosquito *Anopheles funestus* is one of the major malaria vectors in Sub-Saharan Africa. Although it can have a sparse and patchy distribution, this mosquito species is present nearly everywhere across the continent from the savannas of West-Africa, the rainforest of Central Africa, through the dry valleys of East Africa until the Red Island of Madagascar. *Anopheles funestus* breeds in natural and artificial, permanent or semi-permanent water bodies such swamps or rice fields. It is a member of a species group containing at least thirteen species, among which it is the most medically important species. Its prominent role in the transmission of the malaria parasites is due to its close relation to humans, which provide shelters, breeding sites, and blood meals. Although this association makes it highly susceptible to vector campaigns such as indoor residual spraying (IRS) and insecticide treated nets (ITNs), this mosquito species has become resistant to multiple insecticides in many parts of Africa. Therefore, any program aiming at eradicating malaria cannot ignore this species.

At the genetic level, *Anopheles funestus* has been historically neglected in comparison to the members of the *Anopheles gambiae* complex. Multiple studies using genetic markers, such as microsatellites, chromosomal inversions, or DNA sequences have revealed the extraordinary genetic and inversion polymorphism of this species. This genetic richness is likely to underlie its ecological plasticity, its ability to overcome insecticide pressures, and incipient speciation. The first complete genome draft of this mosquito appeared in 2015, originating from a colony derived from wild individuals collected in Mozambique (Fumoz). Later, the quality of the reference genome for Fumoz was improved using long read sequencing from multiple individuals. Here, as part of the *Anopheles* Reference Genomes Project (PRJEB5169), we present a chromosomally complete genome sequence for *Anopheles funestus*, based on a single female specimen from La Lopé, Gabon.

Genome sequence report
The genome was sequenced from a single female *Anopheles funestus* collected from La Lopé, Gabon (-0.187, 11.611). A total of 56-fold coverage in Pacific Biosciences single-molecule long reads (N50 10.684 kb) and 68-fold coverage in 10X Genomics read clouds were generated. Primary assembly contigs were scaffolded with chromosome conformation Hi-C data from a female sibling. Manual assembly curation corrected four missing joins or misjoins, reducing the scaffold number by 0.6%.

The final assembly has a total length of 251 Mb in 330 sequence scaffolds with a scaffold N50 of 84.637 Mb (Table 1). 92.38% of the assembly sequence was assigned to three chromosomal-level scaffolds, representing two autosomes (numbered and oriented against the AfunF3 assembly (9; GCA_003951495.1)), and the X sex chromosome (Figure 1–Figure 4; Table 2). Synteny analysis against the AfunF3 assembly revealed multiple inversions and translocations (Figure 5), correspondence of four largest inversions to known polymorphic inversions in *Anopheles funestus* was revealed based on population genomics and cytogenetics data (Table 3).

The assembly has a BUSCO 5.3.2 completeness of 97.6% using the diptera_odb10 reference set. While not fully phased, the assembly deposited is of one haplotype. Contigs corresponding to the second haplotype have also been deposited.

Methods
Sample acquisition and nucleic acid extraction
*Anopheles funestus* offspring were reared from a wild caught gravid female collected from La Lopé, Gabon (latitude -0.187, longitude 11.611) by Ousman Akone-ella. A single female

<table>
<thead>
<tr>
<th>Table 1. Genome data for <em>Anopheles funestus</em>, idAnoFuneDA-416_04.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Project accession data</strong></td>
</tr>
<tr>
<td>Assembly accession</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>Pacific Biosciences</td>
</tr>
<tr>
<td>10X Genomics Illumina</td>
</tr>
<tr>
<td>Hi-C Illumina</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>

* BUSCO scores based on the diptera_odb10 (3285) set using $BUSCO$ 5.3.2. 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://blobcookiit.genomehubs.org/view/Anopheles%20funestus/dataset/CALSEJ01/busco.
idAnoFuneDA-416_04 was used for Pacific BioSciences and 10x genomics, its sibling female idAnoFuneDA-416_06 was used for Arima Hi-C.

For the high molecular weight (HMW) DNA extraction for *Anopheles* mosquitoes, one whole insect was disrupted by manual grinding with a blue plastic pestle in Qiagen MagAttract lysis buffer and then extracted using the Qiagen MagAttract HMW DNA extraction kit with two minor modifications. These modifications include using half volumes of the kit recommendations due to small sample size (*Anopheles* mosquitoes typically weigh 2–3 mg) and running two elutions of 100 µl each to increase DNA yield. The quality of the DNA was evaluated using an Agilent FemtoPulse to ensure that most DNA molecules were larger than 30 kb, and preferably > 100 kb. Single mosquito extractions ranged in total estimated DNA yield from 192 ng to 800 ng, with an average yield of 500 ng. Low molecular weight DNA was removed from using an 0.8X AMPure XP purification. A small aliquot (<5% of the total volume) of HMW DNA was set aside for 10X Linked Read sequencing and the rest of the DNA was sheared to an average fragment size of 12–20 Kb using a Diagenode Megaruptor 3 at speeds ranging from 27 to 30. Following shearing, samples were cleaned using a SPRI clean up sometimes with a bead ratio intended to help remove fragments below 3 kb. These sheared, cleaned DNAs were then evaluated again on the FemtoPulse and with PicoGreen. The median DNA fragment size was 15 kb and the median concentration of sheared DNA was 200 ng, with samples typically losing about 50% of the original estimated DNA quantity through the process of shearing and purification.
Sheared DNA was purified using AMPure PB beads with a 1.8X ratio of beads to sample to remove the shorter fragments and concentrate the DNA sample. The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer and Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system once more on the sheared and cleaned sample.

For Hi-C samples, a separate sibling whole insect specimen idAnoFuneDA-416_06 was used as input material for the Arima V2 Kit according to the manufacturer’s instructions for animal tissue. This approach of using a sibling was taken in order to enable all material from a single specimen to contribute to the PacBio data generation given we were not always able to meet the minimum suggested guidance of starting with > 300 ng of HMW DNA from a specimen. Samples proceeded to the Illumina library prep stage even if they were suboptimal (too little tissue) going into the Arima reaction.

To assist with annotation, which will be made available through VEuPathDB Vectorbase in due course, RNA was extracted from separate whole unrelated insect specimens.
idAnoFuneDA-146_02, idAnoFuneDA-367_03, and idAnoFuneDA-367_04 using TRIzol, according to the manufacturer’s instructions. RNA was then eluted in 50 µl RNase-free water and its concentration assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit RNA Broad-Range (BR) Assay kit. Analysis of the integrity of the RNA was done using Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay. Samples were not always ideally preserved for RNA, so qualities varied but all were sequenced anyway.

Sequencing
We prepared libraries as per the PacBio procedure and checklist for SMRTbell Libraries using Express TPK 2.0 with low DNA input. Every library was barcoded to support multiplexing. Final library concentrations ranged from 20 ng to 100 ng, and yields were typically only about 25% of the input sheared DNA. Libraries from two specimens were typically multiplexed on a single 8M SMRT Cell. Sequencing complexes were made using Sequencing Primer v4 and DNA Polymerase v2.0. Sequencing was carried out on the Sequel II system with 24 hour run time and 2 hour pre-extension. A 10X Genomics Chromium read cloud sequencing library was also constructed according to the manufacturer’s instructions (this product is no longer available). Only 0.5ng of DNA was used and only 25–50% of the gel emulsion was put forward for library prep due to the small genome size. For Hi-C data generation, following the Arima HiC 2 reaction, samples were processed through Library Preparation using a NEB Next Ultra II DNA Library Prep Kit and sequenced aiming for 100x depth. RNA libraries were created using the directional NEB Ultra II stranded kit. Sequencing was performed by the Scientific Operations core at the Wellcome Sanger Institute on Pacific Biosciences SEQUEL II (HiFi), Illumina NovaSeq 6000 (10X and Hi-C), or Illumina HiSeq 4000 (RNAseq).
Figure 4. Genome assembly of *Anopheles funestus*, idAnoFuneDA-416_04: Hi-C contact map. Hi-C contact map of the idAnoFuneDA-416_04 assembly, visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom. The interactive Hi-C map can be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=ajmC2VieTICjIrBC-4LxgA.

Table 2. Chromosomal pseudomolecules in the genome assembly of *Anopheles funestus*, idAnoFuneDA-416_04.

<table>
<thead>
<tr>
<th>INSDC accession</th>
<th>Chromosome</th>
<th>Size (Mb)</th>
<th>Gaps</th>
</tr>
</thead>
<tbody>
<tr>
<td>OX030923.1</td>
<td>2RL</td>
<td>102.884</td>
<td>3</td>
</tr>
<tr>
<td>OX030924.1</td>
<td>3RL</td>
<td>84.637</td>
<td>4</td>
</tr>
<tr>
<td>OX030925.1</td>
<td>X</td>
<td>22.264</td>
<td>2</td>
</tr>
</tbody>
</table>

Genome assembly
Assembly was carried out with Hiiasm\(^1\); haplotypic duplication was identified and removed with purge_dups\(^4\). One round of polishing was performed by aligning 10X Genomics read data to the assembly with longranger align, calling variants with freebayes\(^5\). The assembly was then scaffolded with Hi-C data\(^6\) using SALSA\(^7\). The assembly was checked for contamination as described previously\(^8\). Manual curation was performed using gEVAL\(^9\), HiGlass\(^10\) and Pretext. The mitochondrial genome was assembled using MitoHiFi\(^11\), which performs annotation using MitoFinder\(^12\). The genome was analysed and BUSCO scores generated within the BlobToolKit environment\(^13\). Synteny analysis was performed with syri v1.6\(^14\) and visualised with plotsr 0.5.3\(^15\). Table 4 contains a list of all software tool versions used, where appropriate.

Figure 5. Synteny between genome assemblies of *Anopheles funestus*, AfunF3 and idAnoFuneDA-416_04. Four largest inversions identified as 2Rh, 3Ra, 3Rb, and 3La (Table 3).
Table 3. Known large-scale inversions between genome assemblies of *Anopheles funestus*, AfunF3 and idAnoFuneDA-416_04 identified by syri. Coordinates given for AfunF3 (ref) and idAnoFuneDA-416_04 (q).

<table>
<thead>
<tr>
<th>Inversion</th>
<th>Chromosome</th>
<th>Start in ref</th>
<th>End in ref</th>
<th>Start in q</th>
<th>End in q</th>
</tr>
</thead>
<tbody>
<tr>
<td>2Rh</td>
<td>2RL</td>
<td>29671756</td>
<td>38731257</td>
<td>29517507</td>
<td>38117412</td>
</tr>
<tr>
<td>3Ra</td>
<td>3RL</td>
<td>1995204</td>
<td>11122881</td>
<td>2428547</td>
<td>12234590</td>
</tr>
<tr>
<td>3Rb</td>
<td>3RL</td>
<td>21099051</td>
<td>33133207</td>
<td>21361107</td>
<td>34095918</td>
</tr>
<tr>
<td>3La</td>
<td>3RL</td>
<td>66185724</td>
<td>86544782</td>
<td>57224763</td>
<td>76848507</td>
</tr>
</tbody>
</table>

Table 4. Software tools used.

<table>
<thead>
<tr>
<th>Software tool</th>
<th>Version</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>hifiasm</td>
<td>0.14</td>
<td>13</td>
</tr>
<tr>
<td>purge_dups</td>
<td>1.2.3</td>
<td>14</td>
</tr>
<tr>
<td>SALSA2</td>
<td>2.2-4c80ac1</td>
<td>17</td>
</tr>
<tr>
<td>longranger</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</td>
<td>15</td>
</tr>
<tr>
<td>MitoHiFi</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>gEVAL</td>
<td>N/A</td>
<td>19</td>
</tr>
<tr>
<td>HiGlass</td>
<td>1.11.6</td>
<td>20</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>23</td>
</tr>
</tbody>
</table>

Ethics/compliance issues
The genetic resources accessed and utilised under this project were done so in accordance with the UK ABS legislation (Nagoya Protocol (Compliance) (Amendment) (EU Exit) Regulations 2018 (SI 2018/1393)) and the national ABS legislation within the country of origin, where applicable.

Data availability

The genome sequence is released openly for reuse. The *Anopheles funestus* genome sequencing initiative is part of the Anopheles Reference Genomes project. All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1.

Author information

References


