DATA NOTE

A chromosomal reference genome sequence for the malaria mosquito, *Anopheles gambiae*, Giles, 1902, Ifakara strain [version 1; peer review: 2 approved]

Tibebu Habtewold[1], Martin Wagah[2], Mgeni Mohamed Tambwe[3], Sarah Moore[3,4], Nikolai Windbichler[1], George Christophides[1], Harriet Johnson[5], Haynes Heaton[6], Joanna Collins[2], Ksenia Krasheninnikova[2], Sarah E. Pelan[2], Damon-Lee B. Pointon[2], Ying Sims[2], James W. Torrance[2], Alan Tracey[2], Marcela Uliano Da Silva[2], Jonathan MD Wood[2], Katharina von Wyschetzki[2], Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective, Shane A. McCarthy[2], Daniel E. Neafsey[7,8], Alex Makunin[2]*, Mara Lawniczak[2]*

[1]Department of Life Sciences, Imperial College London, London, UK
[2]Tree of Life, Wellcome Sanger Institute, Hinxton, UK
[3]Vector Control Product Testing Unit, Ifakara Health institute, Bagamoyo, Tanzania
[4]Vector Biology Unit, Swiss Tropical and Public Health Institute, Bagamoyo, Tanzania
[5]Scientific Operations, Wellcome Sanger Institute, Hinxton, UK
[6]CSSE, Auburn University, Auburn, Alabama, USA
[7]Department of Immunology and Infectious Diseases, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
[8]Infectious Disease and Microbiome Program, Broad Institute, Cambridge, Massachusetts, USA

* Equal contributors

v1 **First published:** 13 Feb 2023, **8**:74
https://doi.org/10.12688/wellcomeopenres.18854.1

Latest published: 13 Feb 2023, **8**:74
https://doi.org/10.12688/wellcomeopenres.18854.1

Abstract

We present a genome assembly from an individual female *Anopheles gambiae* (the malaria mosquito; Arthropoda; Insecta; Diptera; Culicidae), Ifakara strain. The genome sequence is 264 megabases in span. Most 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 gambiae, African malaria mosquito, genome sequence, chromosomal

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 13 Feb 2023	✓ view	✓ view

1. **Alexander W E Franz**, University of Missouri, Columbia, USA
   Zachary Speth, University of Missouri, Columbia, USA
2. **Jason Miller**, Shepherd University, Shepherdstown, USA

This article is included in the Tree of Life gateway.

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding authors: Alex Makunin (am60@sanger.ac.uk), Mara Lawniczak (mara@sanger.ac.uk)

Author roles: Habtewold T: Investigation; **Wagah M**: Visualization; **Tambwe MM**: Investigation; **Moore S**: Investigation; **Windbichler N**: Investigation; **Christophides G**: Investigation; **Johnson H**: Investigation, Methodology; **Heaton H**: Investigation, Methodology; **Collins J**: Data Curation; **Krasheninnikova K**: Investigation, Methodology; **Pelan SE**: Data Curation; **Pointon DLB**: Data Curation; **Sims Y**: Data Curation; **Torrance JW**: Data Curation; **Tracey A**: Data Curation; **Uliano Da Silva M**: Methodology; **Wood JM**: Data Curation; **von Wyschetzki K**: Methodology; **McCarthy SA**: Methodology; **Neafsey DE**: 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 Wellcome through core funding to the Wellcome Sanger Institute (206194, https://doi.org/10.35802/206194), which supports ML. This work was also supported by the Bill and Melinda Gates Foundation (INV-009760 to ML; OPP1158151 to NW and GC). Field collection was supported by an ANR grant (ANR-18-CE35-0002-01 –WILDING). SAM is also supported by 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.

How to cite this article: Habtewold T, Wagah M, Tambwe MM *et al.* **A chromosomal reference genome sequence for the malaria mosquito, *Anopheles gambiae*, Giles, 1902, Ifakara strain [version 1; peer review: 2 approved]** Wellcome Open Research 2023, **8**:74 https://doi.org/10.12688/wellcomeopenres.18854.1

First published: 13 Feb 2023, **8**:74 https://doi.org/10.12688/wellcomeopenres.18854.1

Species taxonomy

Animalia; Arthropoda; Insecta; Diptera; Culicidae; Anophelinae; Anopheles; *Anopheles gambiae*; Giles, 1902 (NCBI txid:7165).

Background

20 years ago, the African malaria mosquito *Anopheles gambiae* became the second insect to have a reference genome[1]. This species is an incredibly important human malaria vector in Africa and the original reference genome ('PEST') is heavily used by a large community studying the biology of this important species. Although the PEST reference has been improved over the years (e.g. 2), the colony has since become extinct and it became clear that it was a mixture of what are known today to be two incipient species: *Anopheles gambiae sensu stricto* (*s.s.* or simply *An. gambiae*) and *Anopheles coluzzii*. Therefore, we sought to create a new *An. gambiae* reference from an extant colony for the large community of users who are working on this species. Technological improvements in recent years mean we can generate reference genomes from single insects using long reads vastly improving the quality of the genome. Here we present a new reference genome for *An. gambiae* s.s., sequenced as part of the Anopheles Reference Genomes Project (PRJEB51690). This genome derives from a single lab-reared female from an extant colony from Tanzania known as the Ifakara strain. This colony is likely to be heterokaryotypic for the 2La inversion, but the primary assembly presented here is 2L+ standard and given colinearity with PEST, is likely to be standard for other common inversions as well. The Ifakara strain has colonies available in Tanzania and the UK and it is available for additional labs by contacting George Christophides. This new reference genome has only 33 gaps across the three chromosomes and at 264 Mb is also 39 Mb larger than the PEST chromosomal assembly (~225 Mb when excluding Ns). This is in comparison to over 6000 gaps in the PEST chromosomes, as well as a bin of contigs containing 27.3 Mb (excluding Ns) of sequences not placed on the three chromosomes. The PEST genome has been an incredibly important genomic resource for the past 20 years to the large community working on both *An. gambiae* and *An. coluzzii*, but there is now an increasing need to differentiate between these two species. The Ifakara strain reference genome will soon have an annotation available via VectorBase, and we encourage studies on *An. gambiae* to make use of this new reference genome instead of the PEST assembly.

Genome sequence report

The genome was sequenced from a single female *An. gambiae* reared in Imperial College London, UK. The Ifakara strain was started from mosquitoes collected in Njage, Tanzania (-8.234, 36.166) in 1996[3]. A total of 54-fold coverage in Pacific Biosciences single-molecule HiFi long reads (N50 10.760 kb) and 77-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 20 missing joins (misjoins) and removed 6 retained haplotigs, reducing the primary assembly size by 1.0% and reducing the scaffold number by 7.8%.

The final assembly has a total length of 264 Mb in 191 sequence scaffolds with a scaffold N50 of 99.150 Mb (Table 1). 92.29% of the assembly sequence was assigned to three chromosomal-level scaffolds, representing two autosomes (numbered and oriented against the AgamP3 assembly ([2]; GCF_000005575.2)), and the X chromosome (Figure 1–Figure 4; Table 2). Synteny analysis against the AgamP3 assembly revealed overall collinearity between the genomes and significant increase in recovery of heterochromatic regions (Figure 5). The total

Table 1. Genome data for *An. gambiae*, idAnoGambNW_F1_1.

Project accession data	
Assembly identifier	idAnoGambNW_F1_1
Species	*Anopheles gambiae*
Specimen	idAnoGambNW-F1_1
NCBI taxonomy ID	7165
BioProject	PRJEB53260
BioSample ID	ERS10527367
Isolate information	female, whole organism
Raw data accessions	
PacificBiosciences SEQUEL II	ERR9439502
10X Genomics Illumina	ERR9356803, ERR9356804, ERR9356805, ERR9356806
Hi-C Illumina	ERR9356802
PolyA RNA-Seq Illumina	ERR9356809, ERR9356810
Genome assembly	
Assembly accession	GCA_943734735
Accession of alternate haplotype	GCA_943734675
Span (Mb)	264.467
Number of contigs	232
Contig N50 length (Mb)	10.625
Number of scaffolds	191
Scaffold N50 length (Mb)	99.150
Longest scaffold (Mb)	118.197
BUSCO* genome score	C:97.3%[S:97.1%,D:0.2%], F:0.7%,M:2.1%,n:3285

* BUSCO scores based on the diptera_odb10 BUSCO 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://blobtoolkit.genomehubs.org/view/idAnoGambNW_F1_1/dataset/CALSDY01.1/busco.

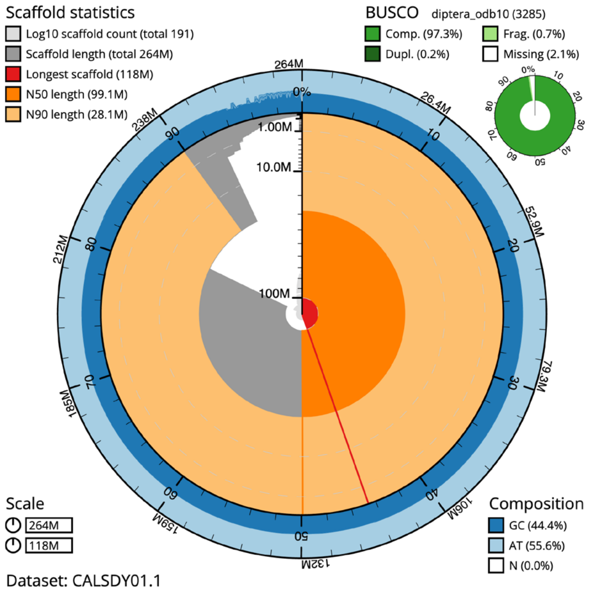

Figure 1. Snail plot summary of assembly statistics for *An. gambiae* assembly idAnoGamb_NW_F1_1. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 264,466,745 bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly (118,196,952 bp, shown in red). Orange and pale-orange arcs show the N50 and N90 chromosome lengths (99,149,756 and 28,097,889 bp), respectively. The pale grey spiral shows the cumulative chromosome count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot. A summary of complete, fragmented, duplicated and missing BUSCO genes in the diptera_odb 10 set is shown in the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idAnoGambNW_F1_1/dataset/CALSDY01.1/snail.

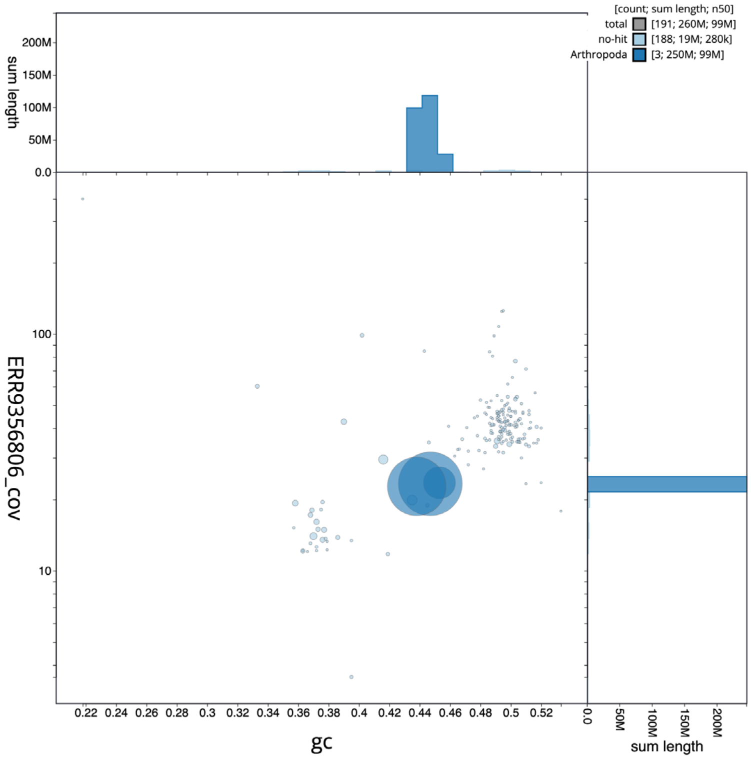

Figure 2. Blob plot of base coverage in a subset of idAnoGambNW_F1_1 10x linked reads against GC proportion for *An. gambiae* assembly idAnoGambNW_F1_1. Chromosomes are coloured by phylum. Circles are sized in proportion to chromosome length. Histograms show the distribution of chromosome length sum along each axis. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idAnoGambNW_F1_1/dataset/CALSDY01.1/blob.

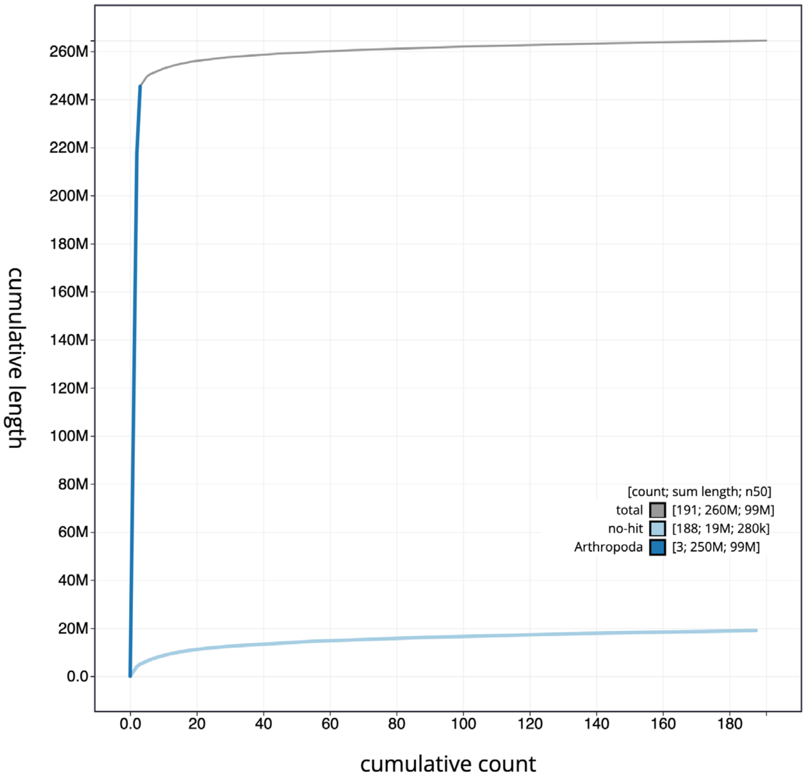

Figure 3. Cumulative chromosome length for *An. gambiae* assembly idAnoGambNW_F1_1. The grey line shows cumulative length for all chromosomes. Coloured lines show cumulative lengths of chromosomes assigned to each phylum using the buscogenes taxrule. The interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/idAnoGambNW_F1_1/dataset/CALSDY01.1/cumulative.

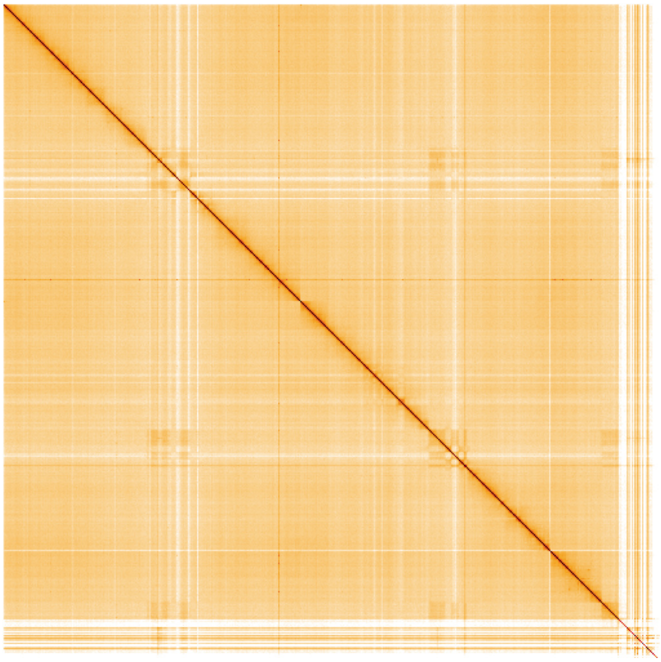

Figure 4. Genome assembly of *An. gambiae*, idAnoGambNW_F1_1: Hi-C contact map. Visualised in HiGlass. Chromosomes are arranged in size order from left to right and top to bottom. The interactive Hi-C map can be viewed at https://genome-note-higlass.tol.sanger.ac.uk/l/?d=MyLlQiYeQHmR7E5i8sjdiw.

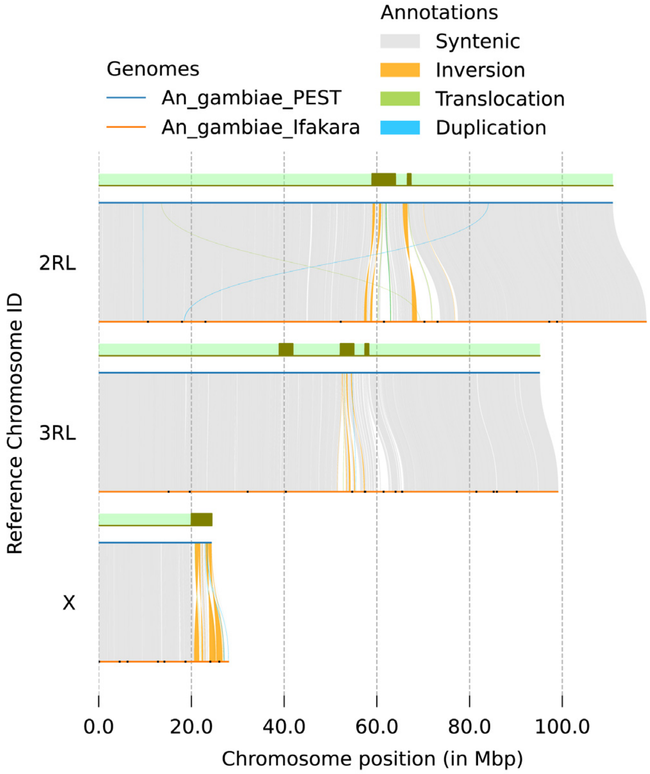

Figure 5. Synteny between genome assemblies of *An. gambiae*, AgamP3 (PEST) and idAnoGambNW_F1_1 (Ifakara). Grey rectangles on green background represent positions of pericentric and intercalary heterochromatin in AgamP3[4]. Remaining gaps in idAnoGambNW_F1_1 indicated with black dots.

Table 2. Chromosomal pseudomolecules in the genome assembly of *An. gambiae*, idAnoGambNW_F1_1.

INSDC accession	Chromosome	Size (Mb)	Count	Gaps
OX030907.1	2RL	118.197	1	9
OX030908.1	3RL	99.150	1	15
OX030909.1	X	28.098	1	9
OX030910.1	MT	0.015	1	0
	X Unlocalised	11.519	161	2
	Unplaced	7.487	26	6

number of assembly gaps across the three chromosomes was reduced dramatically from 6,302 in PEST (AgamP3) to 33 in our assembly (Figure 5, Table 2).

The assembly has a BUSCO 5.3.2[5] completeness of 97.3% 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 gambiae offspring were reared from a lab-reared gravid female of Ifakara strain by Tibebu Habtewold. A single female idAnoGambNW-F1_1 was used for Pacific BioSciences and 10x genomics, and its sibling female idAnoGambNW-F1_3 was used for Arima Hi-C, as described below.

For high molecular weight (HMW) DNA extraction, one whole insect (idAnoGambNW-F1_1) was disrupted by manual grinding with a blue plastic pestle in Qiagen MagAttract lysis buffer and DNA was extracted using the Qiagen MagAttract HMW DNA extraction kit with two minor modifications including halving the volumes recommended by the manufacturer due to small sample size (*Anopheles* mosquitoes typically weigh 2-3 mg) and running two elution steps 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. In general, single *Anopheles* extractions range in total estimated DNA yield from 200 ng to 800 ng, with an average yield of 500 ng. Low molecular weight DNA was removed using an 0.8X AMpure XP purification. A small aliquot (less than ~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. 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 and quality of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer with the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sheared and cleaned sample on the FemtoPulse system once more. The median DNA fragment size for *Anopheles* mosquitoes 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.

For Hi-C data generation, a separate sibling mosquito specimen (idAnoGambNW-F1_3) 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 sibling insect specimens (idAnoGambNW-F1_9 and idAnoGambNW-F1_10) using TRIzol, according to the manufacturer instructions. RNA was then eluted in 50 μl RNAse-free water, and its concentration was 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 a 24-hour run time and a 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 0.5 ng 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 V2 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).

Genome assembly

Assembly was carried out with Hifiasm[6]; haplotypic duplications were identified and removed with purge_dups[7]. One round of polishing was performed by aligning 10X Genomics read data to the assembly with longranger align, calling variants with freebayes[8]. The assembly was then scaffolded with Hi-C data[9] using SALSA2[10]. The assembly was checked for contamination as described previously[11]. Manual curation was performed using gEVAL[12], HiGlass[13] and Pretext[14]. The mitochondrial genome was assembled using MitoHiFi[15], which performs annotation using MitoFinder[16]. The genome was analysed and BUSCO scores were generated within the BlobToolKit environment[17]. Synteny analysis was performed with syri[18] and visualised with plotsr[19]. Table 3 contains a list of all software tool versions used, where appropriate.

Table 3. Software tools used.

Software tool	Version	Source
hifiasm	0.14	6
purge_dups	1.2.3	7
SALSA2	2.2-4c80ac1	10
longranger align	2.2.2	20
freebayes	1.3.1	8
MitoHiFi	2	15
gEVAL	N/A	12
HiGlass	1.11.6	13
PretextView	0.1.x	14
BlobToolKit	3.4.0	17
BUSCO	5.3.2	5
syri	1.6	18
plotsr	0.5.3	19

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

NCBI BioProject: *Anopheles gambiae* genome assembly, idAnoGambNW_F1_1. Accession number PRJEB53260; https://identifiers.org/bioproject/PRJEB53260[21].

The genome sequence is released openly for reuse. The *Anopheles gambiae* genome sequencing initiative is part of the Anopheles Reference Genomes project PRJEB51690. 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

Members of Wellcome Sanger Institute Scientific Operations: DNA Pipelines collective are listed here: https://doi.org/10.5281/zenodo.4790455.

References

1. Holt RA, Subramanian GM, Halpern A, *et al.*: **The genome sequence of the malaria mosquito *Anopheles gambiae*.** *Science.* 2002; **298**(5591): 129–149.
PubMed Abstract | **Publisher Full Text**

2. Sharakhova MV, Hammond MP, Lobo NF, *et al.*: **Update of the *Anopheles gambiae* PEST genome assembly.** *Genome Biol.* 2007; **8**(1): R5.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

3. Huho BJ, Ng' habi KR, Killeen GF, *et al.*: **Nature beats nurture: a case study of the physiological fitness of free-living and laboratory-reared male *Anopheles gambiae* s.l.** *J Exp Biol.* 2007; **210**(Pt 16): 2939–2947.
PubMed Abstract | **Publisher Full Text**

4. Sharakhova MV, George P, Brusentsova IV, *et al.*: **Genome mapping and characterization of the *Anopheles gambiae* heterochromatin.** *BMC Genomics.* 2010; **11**: 459.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

5. Simão FA, Waterhouse RM, Ioannidis P, *et al.*: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics.* 2015; **31**(19): 3210–3212.
PubMed Abstract | **Publisher Full Text**

6. Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

7. Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

8. Garrison E, Marth G: **Haplotype-based variant detection from short-read sequencing.** *arXiv [q-bio.GN].* 2012.
Publisher Full Text

9. Rao SS, Huntley MH, Durand NC, *et al.*: **A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | **Publisher Full Text**| **Free Full Text**

10. 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**

11. Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

12. Chow W, Brugger K, Caccamo M, *et al.*: **gEVAL - a web-based browser for evaluating genome assemblies.** *Bioinformatics.* 2016; **32**(16): 2508–2510.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

13. 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**

14. PretextView: **OpenGL Powered Pretext Contact Map Viewer.** Github.
Reference Source

15. Uliano-Silva M, Nunes JGF, Krasheninnikova K, *et al.*: **marcelauliano/MitoHiFi: mitohifi_v2.0.** 2021.
Publisher Full Text

16. Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

17. Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – Interactive Quality Assessment of Genome Assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

18. Goel M, Sun H, Jiao WB, *et al.*: **SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies.** *Genome Biol.* 2019; **20**(1): 277.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

19. Goel M, Schneeberger K: **plotsr: visualizing structural similarities and rearrangements between multiple genomes.** *Bioinformatics.* 2022; **38**(10): 2922–2926.
PubMed Abstract | **Publisher Full Text** | **Free Full Text**

20. **Long Ranger BASIC and ALIGN Pipelines -Software -Genome & Exome-Official 10x Genomics Support.** [cited 16 Dec 2022].
Reference Source

21. Wellcome Sanger Institute: **Anopheles gambiae genome assembly, idAnoGambNW_ F1_1, NCBI BioProject.** [dataset], 2022; accession number PRJEB53260.

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 1

Reviewer Report 06 July 2023

https://doi.org/10.21956/wellcomeopenres.20906.r61374

✓ **Jason Miller**
Computers, Mathematics, Engineering, Shepherd University, Shepherdstown, West Virginia, USA

This data note describes an assembly of the malaria mosquito *Anopheles gambiae*. The new assembly improves upon, and will likely replace, the existing reference. Unlike the prior reference, the new sequence is based on DNA from a single individual. Despite the small quantity of available DNA, the authors heroically applied hifi long read, 10X composite read, and HiC technology. The paper does not include biological inference or even gene annotation beyond BUSCO analysis, but this is in accordance with the project goals to release assemblies quickly, and the journal's instructions for data notes.

Minor Revisions Suggested:

The paper should clarify which assembly statistics were generated after the manual curation exclusively. For example, if the manual removal of 6 haplotigs was informed by pre-curation BUSCO duplicated gene analysis, then the post-curation BUSCO results should be understood to be a curation objective that was met, and not an independent validation.

The background should clear up the reference nomenclature and history. In this version, the background refers to PEST exclusively, but the next section refers to AgamP3, and later to "PEST (AgamP3)". The background should not say PEST was "incredibly important" (twice). It would be more convincing to cite discoveries enabled by PEST. The background should clarify the sources of its claims that "the colony has since become extinct" and that "it became clear that it was a mixture" of species. The authors should cite or explain "the 2La inversion" and why they believe "the colony is likely to be heterokaryotypic."

The paper could briefly compare its methods and results to those of the Anopheles 16 genomes project (Neafsey 2013)[1], the Anopheles 1000 genomes project (Genome Research 2020)[2], and other chromosome-scale mosquito assemblies (e.g. Ghurye et al 2019[3] and Ayala et al 2022[4]). INSDC could be referenced (Brunak 2002[5]).

Figure 5 and Table 2 are clear, helpful, and impressive. Figures 1 to 4 are unhelpful just because the whole genome is in 4 scaffolds. For example, the complicated snail plot is just a polar chart of the four scaffold lengths from Table 2. Figure 4 shows nearly uniform background noise, which looks good, but it is hard to interpret quantitatively without a color legend or a comparative example. It seems this paper follows the manuscript template used by other products of this project. I would not remove the figures, but I would add text to tell the reader what the figures show. In this version, the text links to the figures without explanation and the captions give only technical details.

References

1. Neafsey DE, Christophides GK, Collins FH, Emrich SJ, et al.: The evolution of the Anopheles 16 genomes project.*G3 (Bethesda)*. 2013; **3** (7): 1191-4 PubMed Abstract | Publisher Full Text
2. Anopheles gambiae 1000 Genomes Consortium: Genome variation and population structure among 1142 mosquitoes of the African malaria vector species Anopheles gambiae and Anopheles coluzzii.*Genome Res*. 2020; **30** (10): 1533-1546 PubMed Abstract | Publisher Full Text
3. Ghurye J, Koren S, Small ST, Redmond S, et al.: A chromosome-scale assembly of the major African malaria vector Anopheles funestus.*Gigascience*. 2019; **8** (6). PubMed Abstract | Publisher Full Text
4. Ayala D, Akone-Ella O, Kengne P, Johnson H, et al.: The genome sequence of the malaria mosquito, Anopheles funestus, Giles, 1900. *Wellcome Open Research*. 2022; **7**. Publisher Full Text
5. Brunak S, Danchin A, Hattori M, Nakamura H, et al.: Nucleotide Sequence Database Policies. *Science*. 2002; **298** (5597). Publisher Full 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?

Yes

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, genomics, transcriptomics, machine learning.

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 12 June 2023

https://doi.org/10.21956/wellcomeopenres.20906.r58446

✓ **Alexander W E Franz**
University of Missouri, Columbia, Missouri, USA
Zachary Speth
University of Missouri, Columbia, MO, USA

Presented is the highest quality genome assembly for *An. gambiae s.s.* generated to date. With the application of three long read sequencing methods, Habtewold *et al.* constructed chromosome length assemblies with coverage of the two *An. gambiae* autosomes 2RL and 3RL, as well as the X chromosome and the mitochondrial chromosome. This improved reference genome is a highly useful resource to any researcher currently studying *An. gambiae*.

The most advanced/appropriate sequencing methods were used in this work. The overall coverage level of >50-fold (single-molecule HIFi long reads) and >70-fold in 10xGenomics reads looks sufficient to reliably distinguish SNPs between haplotypes. The genome data are available at NCBI BioProject. The data are accessible as checked for the X chromosome (as an example). Figures 2, 3, and 4 are not easy to understand for a non-specialist.

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: mosquito transgenesis, transcriptome/proteome analysis, single-cell transcriptomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.