BMC Research Notes

DATA NOTE Open Access

Data of the study: *De Novo* genome assembly, annotation, and characterization of chemosensory genes in the camel ked (*Hippobosca camelina*)

Fredrick Kebaso[1,3], Caleb Kibet[1], JohnMark O. Makwatta[1], Dennis Getange[1], Souleymane Diallo[1], Suhaila Hashim[3], Jean-Baka Domelevo Entfellner[2] and Daniel Masiga[1*]

Abstract

Objectives *Hippobosca camelina* (camel ked) is an obligate hematophagous ectoparasite that infests camels. Hematophagy inflicts painful bites, leading to myiasis, anaemia and pathogen transmission. The genome assembly for this species is currently unavailable, limiting our understanding of its genetics, particularly its chemosensory system. Therefore, this study focused on generating whole genome assembly and characterizing the chemosensory genes to aid in molecular studies of the camel ked.

Data Description Genomic DNA was extracted from the thoracic muscles of both the male and female camel keds. A total of 11.8 Giga base pairs for the male ked and 13.6 Giga base pairs for the female ked, were sequenced using Oxford Nanopore technology. Reads were then assembled into a 133.5 Mega base pairs genome for the male ked with an N50 of 419.6 Kb and 135.6 Mega base pairs for the female ked with an N50 of 1.2 Mega base pairs. Both genomes had a BUSCO (Genome mode) completion rate above 97% with the dipteran lineage datasets. Active Chemosensory genes recovered included 4 Chemosensory specific proteins (CSPs), 18 Ionotropic receptors (IRs), 7 Gustatory receptors (GRs), 5 Odorant receptors (ORs), 9 Odorant binding proteins (OBPs) and 1 Sensory Neuron Membrane Protein (SNMP). This information will lead to a better understanding of the genomic structure and functional chemosensory genes of the keds.

Keywords Hippobosca camelina, Genome assembly, Chemosensory genes, Oxford Nanopore sequencing, Ectoparasite genomics

*Correspondence:
Daniel Masiga
dmasiga@icipe.org
[1]International Centre of Insect Physiology and Ecology (ICIPE), Nairobi, Kenya
[2]International Livestock Research Institute (ILRI), Nairobi, Kenya
[3]Pwani University, Kilifi, Kenya

Objectives

Hippoboscids (keds or flies) are obligate, blood-sucking ectoparasites that infest birds, mammals, and rarely humans. They transmit pathogens such as *Anaplasma camelii* [1–3]. Keds require approximately 1.5 mg of blood in a span of 6 hours, feeding at an interval of 13 min [4]. As such, heavy infestations would lead to *anaemia* if left unchecked. The frequent feeding intervals result in prurition, prompting the host to scratch against rough surfaces. The scratches lead to the development of wounds, creating a conducive environment for myiasis, a condition where larvae infest the host's wounds. Myiasis adds to the economic burden on farmers in increased production costs and economic losses when livestock succumb to the adverse effects of this infestation.

Control strategies for vector-borne diseases (VBDs) in dipterans closely related to *Hippobosca*, *such as Glossina*, have exploited the chemosensory system that plays a vital role in their ability to locate hosts for feeding, mates for reproduction, and predator evasion, contributing to their successful ecological establishment [5]. A genome assembly for the *Hippobosca* species is currently unavailable in the genome database, which limits understanding of its genetics, particularly the chemosensory system. Therefore, this study focused on generating whole genome assemblies for *Hippobosca camelina* and describing their chemosensory system genes.

Data description

The male and female keds were collected from Laisamis area in Marsabit County, Kenya. The samples were kept in perforated water bottles to guarantee oxygenation. They were shipped to the Molecular Biology and Bioinformatics Unit (MBBU) laboratories at the International Centre of Insect Physiology and Ecology (ICIPE), Kenya and stored in a -80 C freezer before processing. DNA was extracted from the thoracic muscle (one male and one female fly) using the protein precipitate solution (PPS) method and eluted using 40uL of elution buffer from the Qiagen DNA extraction kit. Nucleic acids were run on a 1% agarose gel to check their integrity. The sequencing library was prepared using the Ligation Sequencing Kit (SQK-LSK109) and sequencing was performed using the Oxford Nanopore MinION Mk1B. Base calling was performed using Dorado version 7.2.13.

The female-ked genomic DNA resulted in 15,935,230 sequences with a read N50 of 1,217 base pairs, an average quality score of 9.4 and a GC percentage of 32.77. The male ked genomic DNA had 17,538,802 sequences, a read N50 of 833 base pairs, an average quality score of 10.11 and a GC percentage of 33.32. Reads were assembled using Flye version 2.9.4 [6]. For the male ked genome, Illumina (Nextseq550) reads from a previous study were mapped to the *de novo* genome, with 99% mapping to the genome. These reads were used to polish the genome using Pilon version 1.24 [7]. The male ked had a total genome size of 133.5 Mega base pairs (133,470,331 base pairs), an N50 of 419.6 kb, 2,318 contigs, a genome coverage of 90x, a GC percentage of 33.5, the largest contig was 2.2 Mb and the shortest contig 220 base pairs. Genome quality was assessed usingcompleasm version 0.2.9 [8]. BUSCO (proteome mode) completion rates (single copy orthologs) were over 94% in reference to the Diptera_odb10 lineage datasets for both genomes, (Male ked: Single copy genes(S) 94.70%; Duplicated genes (D) 0.49%; Fragmented genes subclass 1 (F) 0.85%; Fragmented genes subclass 2 (I) 0.00%, Missing genes (M) 3.96% and Female ked: Single copy genes(S) 95.83%; Duplicated genes (D) 0.30%; Fragmented genes subclass 1 (F) 0.30%; Fragmented genes subclass 2 (I) 0.00%, Missing genes (M) 3.5%). The female ked genome was 135.6 Mega base pairs (135,596,069 base pairs) with an N50 of 1.2 Mega base pairs, 2,182 contigs, a GC percentage of 33.5, a genome coverage of 102x, the largest contig is 5.5 Mb, and the shortest contig had 209 base pairs. The assembled genome quality was assessed using compleasm version 0.2.9 [8]. Braker version 3.0.8 [9–19] was used to perform automated gene prediction on the genomes. Using chemosensory homologs from the closest species including *Anopheles gambiae*, *Glossina morsitans morsitans*, *Glossina fuscipes*, *Glossina brevipalpis*, and *Drosophila* Orthofinder version 2.5 [20] classified the predicted genes into different orthogroups among them chemosensory genes orthogroups. Further filtering and functional analysis including conserved protein domain validation via NCBI's Conserved Domain Database (CDD) [21] BLAST searches, amino acid alignment using MUSCLE5 [22] to verify the presence of key residues required for an active chemosensory gene, and phylogenetic analysis with RaxML [23] validated, 4 Chemosensory Specific Proteins (CSPs), 18 Ionotropic Receptors (IRs), 7 Gustatory Receptors (GRs), 5 Odorant Receptors (ORs), 9 Odorant Binding Proteins (OBPs), and 1 Sensory Neuron Membrane Protein (SNMP) for the camel ked genome.

The table summarizes the data sets generated and used in the study, including genome assemblies, raw genomic DNA sequences, and annotated chemosensory gene accessions. Each data set is described by its label, name, file type, repository, and unique identifier (DOI or accession number).

Limitations

The genome assembly contains many contigs that can be further scaffolded to the chromosomal level using Hi-C, bionano or other techniques.

Table 1 Overview of data files and data sets associated with the study

Labels	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Data Set 1	*Hippobosca camelina* male assembled genome	Fasta files (.fna)	NCBI GenBank accession number: GCA_041146595.1(http://identifiers.org/insdc.gca:GCA_041146595.1)(24).
Data Set 2	*Hippobosca camelina* female assembled genome	Fasta files (.fna)	NCBI GenBank accession number: GCA_041146635.1 (http://identifiers.org/insdc.gca:GCA_041146635.1)(25)
Data set 3	*Hippobosca camelina* female Nanopore Raw genomic DNA sequences	Fastq files (.fastq.gz)	*Sequence Read Archive from NCBI repository, accession number*: SRX25529404 (http://identifiers.org/insdc.sra:SRX25529404)(26)
Data set 4	*Hippobosca camelina* female Nanopore raw genomic DNA sequences	Fastq files (.fastq.gz)	*Sequence Read Archive from NCBI repository, accession number*: SRX25529403 (http://identifiers.org/insdc.sra:SRX25529403)(27)
Data set 5	NCBI GenBank gene accessions for the annotated chemosensory genes	Text file (.pdf)	*Figshare* (https://doi.org/10.6084/m9.figshare.29135213)(28)

Abbreviations

DNA	Deoxyribonucleic Acid
VBD	Vector Borne Diseases
ICIPE	International Centre of Insect Physiology and Ecology
MBBU	Molecular Biology and Bioinformatics Unit
PPS	Protein Precipitation Solution
BUSCO	Benchmarking Universal Single-Copy Orthologs
NCBI	National Center for Biotechnology Information
SRA	Sequence Read Archive

Author contributions

F.K. performed data analysis and wrote the manuscript. C.K., J.B., S.D., S.H., and D.M. reviewed the manuscript. J.M., and D.G. handled sample collection, DNA extraction, and manuscript review. All authors contributed to and approved the final manuscript.

Funding

This study was funded by the National Institutes of Health (NIH) through Eastern Africa Network for Bioinformatics Training - (EANBiT), grant number 5U2RTW010677, Eneza Data Science: Enhancing Data Science Capability and Tools for Health in East Africa, Award number 1UE5TW012539, German Federal Ministry for Economic Cooperation and Development (BMZ) commissioned and administered through the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) Fund for International Agricultural Research (FIA), grant number 81235250.

Data availability

The data described in this data note can be freely and openly accessed from the NCBI GenBank using the accessions given below: (i) ***Hippobosca camelina*** female genome: GCA_041146595.1 ([http://identifiers.org/insdc.gca: GCA_041146595.1] (http:/identifiers.org/insdc.gca:GCA_041146595.1) [24]. (ii) ***Hippobosca camelina*** male genome: GCA_041146635.1 ([http://identifiers.org/insdc.gca: GCA_041146635.1] (http:/identifiers.org/insdc.gca:GCA_041146635.1) [25].The raw reads can be retrieved from Sequence Read Archive (SRA) and can be accessed using the below SRA IDs: (i) ***Hippobosca camelina*** female genomic reads: SRA ID – SRX25529404 ([http://identifiers.org/insdc.sra: SRX25529404] (http:/identifiers.org/insdc.sra:SRX25529404) [26]. (i) ***Hippobosca camelina*** male genome: SRA ID – SRX25529403 ([http://identifiers.org/insdc.sra: SRX25529403] (http:/identifiers.org/insdc.sra: SRX25529403) [27]. See Table 1 for details and links to the data.

Declarations

Ethics approval and consent to participate

The study protocol was reviewed and approved by the Pwani University Ethics Review Board (Approval Number: ERC/EXT/002/2020), accredited by the National Commission for Science, Technology and Innovation (NACOSTI), Kenya. Additionally, all animal-related procedures were approved by the International Centre of Insect Physiology and Ecology's Institutional Animal Care and Use Committee (IACUC) under reference number:IcipeA CUC2018-003–2023.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 28 April 2025 / Accepted: 23 October 2025
Published online: 29 December 2025

References

1. Kidambasi KO, Masiga DK, Villinger J, Carrington M, Bargul JL. Detection of blood pathogens in camels and their associated ectoparasitic camel biting Keds, Hippobosca camelina: the potential application of Keds in xenodiagnosis of camel haemopathogens. AAS Open Res. 2020;2(May):164.
2. Bargul JL, Kidambasi KO, Getahun MN, Villinger J, Copeland RS, Muema JM et al. Transmission of 'candidatus anaplasma camelii' to mice and rabbits by camel-specific keds, hippobosca camelina. PLoS Negl Trop Dis [Internet]. 2021;15(8):1–20. Available from: https://doi.org/10.1371/journal.pntd.0009671
3. Bezerra-Santos MA, Otranto D. Keds, the enigmatic flies and their role as vectors of pathogens. Acta Trop [Internet]. 2020;209(April):105521. Available from: https://doi.org/10.1016/j.actatropica.2020.105521
4. Rani PAMA, Coleman GT, Irwin PJ, Traub RJ. Hippobosca longipennis - A potential intermediate host of a species of Acanthocheilonema in dogs in northern India. Parasit Vectors [Internet]. 2011;4(1):143. Available from: http://www.parasitesandvectors.com/content/4/1/143
5. Malele II, Ouma JO, Nyingilili HS, Kitwika WA, Malulu DJ, Magwisha HB et al. Comparative performance of traps in catching Tsetse flies (Diptera: Glossinidae) in Tanzania. Onderstepoort J Vet Res. 2016;83(1).
6. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6.
7. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE. 2014;9(11).
8. Huang N, Li H. Compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics. 2023;39(10).
9. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32(5):767–9.
10. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP + and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 2021;3(1).
11. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. Methods in molecular biology. Humana Press Inc.; 2019. pp. 65–95.
12. Brůna T, Lomsadze A, Borodovsky M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom

Bioinform [Internet]. 2020 Jun 1 [cited 2025 Mar 6];2(2). Available from: https://pubmed.ncbi.nlm.nih.gov/32440658/
13. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res [Internet]. 2005 [cited 2025 Mar 6];33(20):6494–506. Available from: https://pubmed.ncbi.nlm.nih.gov/16314312/
14. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods [Internet]. 2015 Jan 1 [cited 2025 Mar 6];12(1):59–60. Available from: https://pubmed.ncbi.nlm.nih.gov/25402007/
15. Gotoh O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res [Internet]. 2008 May [cited 2025 Mar 6];36(8):2630–8. Available from: https://pubmed.ncbi.nlm.nih.gov/18344523/
16. Iwata H, Gotoh O. Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. Nucleic Acids Res [Internet]. 2012 Nov [cited 2025 Mar 6];40(20). Available from: https://pubmed.ncbi.nlm.nih.gov/22848105/
17. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de Novo gene finding. Bioinformatics. 2008;24(5):637–44.
18. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;7.
19. Gabriel L, Hoff KJ, Brůna T, Borodovsky M, Stanke M. TSEBRA: transcript selector for BRAKER. BMC Bioinformatics [Internet]. 2021 Dec 1 [cited 2025 Mar 6];22(1):1–12. Available from: https://bmcbioinformatics.biomedcentral.com/articles/https://doi.org/10.1186/s12859-021-04482-0
20. Emms DM, Kelly S, OrthoFinder. Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1).
21. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48(D1):D265–8.
22. Edgar RC. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. Nat Commun. 2022;13(1).
23. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
24. Kebaso F, Kibet C, Makwatta JO, Getange D, Domelevo Entfellner JB, Masiga D. Data of the study: de Novo genome assembly, annotation, and characterization of chemosensory genes in the camel Ked (Hippobosca camelina). GenBank. 2025. http://identifiers.org/insdc.gca:GCA_041146595.1
25. Kebaso F, Kibet C, Makwatta JO, Getange D, Domelevo Entfellner JB, Masiga D. Data of the study: de Novo genome assembly, annotation, and characterization of chemosensory genes in the camel Ked (Hippobosca camelina). GenBank. 2025. http://identifiers.org/insdc.gca:GCA_041146635.1
26. Kebaso F, Kibet C, Makwatta JO, Getange D, Domelevo Entfellner JB, Masiga D. Data of the study: de Novo genome assembly, annotation, and characterization of chemosensory genes in the camel Ked (Hippobosca camelina). NCBI Sequence Read Archive. 2025. http://identifiers.org/insdc.sra:SRX25529404
27. Kebaso F, Kibet C, Makwatta JO, Getange D, Domelevo Entfellner JB, Masiga D. Data of the study: de Novo genome assembly, annotation, and characterization of chemosensory genes in the camel Ked (Hippobosca camelina). NCBI Sequence Read Archive. 2025. http://identifiers.org/insdc.sra:SRX25529403
28. Kebaso F, Kibet C, Makwatta JO, Getange D, Domelevo Entfellner JB, Masiga D. Data of the study: de Novo genome assembly, annotation, and characterization of chemosensory genes in the camel Ked (Hippobosca camelina). 2025. Figshare https://doi.org/10.6084/m9.figshare.29135213

Publisher's note