data

MDPI

Data Descriptor

First Draft Genome Assembly of Tropical Bed Bug, *Cimex hemipterus* (F.)

Li Lim [1] and Abdul Hafiz Ab Majid [1,2,*]

[1] Household & Structural Urban Entomology Laboratory, Vector Control Research Unit, School of Biological Sciences, Universiti Sains Malaysia, George Town 11800, Malaysia; limli110376.ll@gmail.com
[2] Centre for Insect Systematics, Faculty of Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi 43600, Malaysia
* Correspondence: abdhafiz@usm.my

Abstract: *Cimex hemipterus*, a blood-feeding ectoparasite commonly found in tropical regions, is a notorious household pest. The draft genome assembly of *C. hemipterus* is presented in this study, generated using SPAdes software with Illumina short reads. The obtained genome size was 388.66 Mb with a contig N50 size of 3503 bp. BUSCO assessment indicated that 96.71% of the expected Insecta lineage genes were complete in the genome assembly. Annotation of the *C. hemipterus* genome assembly identified 2.88% of repetitive sequences and 17,254 protein-coding genes. Functional annotation showed that most gene families are involved in cellular processes and signaling. This first *C. hemipterus* genome will be helpful in further understanding the bed bug genetics and evolution, while the annotated genome may also help in devising new strategies in bed bug management.

Dataset: The raw genome sequencing data of Illumina HiSeq were deposited in NCBI in FASTQ format with BioSample accession number SAMN18780126 under BioProject PRJNA722579. The assembled data and its annotation files are available in figshare repository (DOI: 10.6084/m9.figshare.16815364).

Keywords: *Cimex hemipterus*; tropical bed bug; draft genome; Illumina sequencing

Citation: Lim, L.; Ab Majid, A.H. First Draft Genome Assembly of Tropical Bed Bug, *Cimex hemipterus* (F.). *Data* **2022**, *7*, 101. https://doi.org/10.3390/data7070101

Received: 24 March 2022
Accepted: 8 June 2022
Published: 21 July 2022

1. Summary

Cimex hemipterus is s a wingless, red-brown, human blood-feeding insect in the Cimicidae family [1]. It is also known as the tropical bed bug due to the fact that this species is typically found in the tropical regions of Southeast Asia, Africa, and South America [2]. Nevertheless, this species may have the ability to expand outside the tropical range, since it has also been reported in temperate regions, including Florida [3], Italy [4], and Paris [5].

Bed bugs develop through five nymphal instars before reaching adulthood, with each life stage requiring a blood meal to molt to the next instar [6]. Thus, bed bug infestations could happen almost anywhere, such as dormitories, residential houses, hotels, and airports, as long as they can contact their hosts [5,7].

Although there is no evidence to prove that they are a vector, bed bug bites can cause skin reactions, such as itching, wheals, and lesions to the majority of bitten individuals and may even lead to anemia and iron inadequacies with heavy bed bug infestations [1,8–10]. Besides physiological responses, the victim can also be affected psychologically, resulting in anxiety, sleep deprivation, and fatigue [11].

Bed bug infestation of human habitats has increased drastically over the last two decades due to the difficulty in controlling bed bugs and human-mediated spreading [5,7,8,12]. This has created a need for renewed research on tropical bed bugs, such as at the molecular level, to further understand the crucial molecular biology processes, such as gene expression and transcription regulation, in order to find a new strategy for pest control. However, this goal

could be impeded without the reference genome. Thus, this is the first study that reports the draft genome and annotation set of *C. hemipterus*.

2. Data Description

2.1. Genome Assembly

The Illumina sequencing data of *Cimex hemipterus* were assembled, and an assembly with a total length of 388.66 Mb and a contig N50 of 3503 bp was yielded (Table 1).

Table 1. Statistics of the *C. hemipterus* genome assembly.

Assembly Statistics	SPAdes
No. scaffolds	166,338
No. scaffolds (≥1000 bp)	113,017
No. scaffolds (≥5000 bp)	16,539
No. scaffolds (≥10,000 bp)	3013
No. scaffolds (≥25,000 bp)	136
No. scaffolds (≥50,000 bp)	10
Total length (Mb)	388.66
N50	3503
N75	1815
L50	30,829
L75	69,477
GC (%)	34.80

BUSCO [13] assessment of the genome indicated that 96.71% of the gene sets were completed (including single and duplicated gene sets), while the rest were either fragmented (2.19%) or missing (1.10%). The data set for *C. hemipterus* was concluded to be sufficiently comprehensive for further downstream analyses.

2.2. Repetitive Elements Identification and Annotation

The identified repetitive elements, including short interspersed nuclear elements (SINEs, 0.09%), long interspersed nuclear elements (LINEs, 0.09%), long terminal repeats (LTR, 0.06%), DNA transposons (0.14%), small RNA (0.23%), satellites (0.03%), simple repeats (1.78%), and low complexity (0.46%), representing 2.88% of the *C. hemipterus* genome assembly, are shown in Table 2.

Table 2. Statistics of repetitive elements identified in the *C. hemipterus* genome assembly.

Category	No. of Elements *	Total (bp)	% of Genome
SINEs	4656	369,602	0.09
LINEs	2294	370,113	0.09
LTR	1062	242,760	0.06
DNA transposons	2493	602,079	0.14
Unclassified	47	4431	0.00
Small RNA	14,718	942,706	0.23
Satellites	230	120,016	0.03
Simple repeats	160,826	7,408,775	1.78
Low complexity	33,694	1,928,180	0.46
Total			2.88

* Most repeats fragmented by insertions or deletions have been counted as one element.

2.3. Protein-Coding Gene Annotation

A total of 17,254 protein-coding genes were estimated from the repeat-elements masked assembly, and a total of 11,431 genes were annotated against the Eukaryotic Orthologous Groups (KOG) database. Analysis showed that most of the genes were associated with cellular processes and signaling (33.0%) followed by metabolism (25.0%), and information storage and processing (24.0%), and 18% of the genes were poorly characterized (Table 3).

Table 3. Genes with putative functions in the tropical bed bug based on the KOG database.

Category	Percentage (%)	
i. Information storage and processing		
a. Translation, ribosomal structure, and biogenesis	8.0	
b. RNA processing and modification	3.0	
c. Transcription	7.0	24.0
d. Replication, recombination, and repair	4.0	
e. Chromatin structure and dynamics	2.0	
ii. Cellular processes and signaling		
a. Cell cycle control, cell division, chromosome partitioning	2.0	
b. Nuclear structure	0.4	
c. Defense mechanisms	0.5	
d. Signal transduction mechanisms	12.0	
e. Cell wall/membrane/envelope biogenesis	1.0	33.0
f. Cell motility	0.1	
g. Cytoskeleton	3.0	
h. Extracellular structures	1.0	
i. Intracellular trafficking, secretion, and vesicular transport	4.0	
j. Post-translational modification, protein turnover, chaperones	9.0	
iii. Metabolism		
a. Energy production and conversion	5.0	
b. Carbohydrate transport and metabolism	3.0	
c. Amino acid transport and metabolism	7.0	
d. Nucleotide transport and metabolism	2.0	25.0
e. Coenzyme transport and metabolism	2.0	
f. Lipid transport and metabolism	3.0	
g. Inorganic ion transport and metabolism	2.0	
h. Secondary metabolites biosynthesis, transport, and catabolism	1.0	
iv. Poorly characterized		
a. General function prediction only	13.0	18.0
b. Function unknown	5.0	

3. Methods

3.1. DNA Extraction, Library Construction, and Sequencing

A lab-reared specimen was used in this study. The bed bug samples were first collected in 2014 from cushioned seats in the waiting area in Kuala Lumpur International Airport (KLIA), Malaysia, and bred in the laboratory subsequently [7]. Before DNA extraction, one adult male bed bug was surface sterilized with 70% ethanol and rinsed with sterile distilled water. The whole organism was then crushed into pieces using a micro-pestle, and the DNA was extracted using the HiYield Genomic DNA isolation kit (Real Biotech Corporation, Taiwan) according to the manufacturer's instructions. Illumina DNA paired-end (PE) libraries were constructed according to the standard protocol provided by Illumina (San Diego, CA, USA), with short-insert sizes (150 bp). Sequencing was then performed on the HiSeq 2000 platform (Illumina, San Diego, CA, USA), and a total of 9 Gb of Illumina reads were produced.

3.2. Genome Assembly and Data Analysis

The following workflow was performed on Linux system Ubuntu (64-bit) with 6 CPUs processor and 13,000 MB base memory. Before downstream analyses, the quality of all the Illumina reads was estimated using FastQC [14]. The adapter-contaminated and low-quality reads (<Q30) were removed using Trimmomatic (ver0.36) [15], yielding a total of 8.66 Gb of clean Illumina reads. The paired-end reads were merged using FLASH (ver1.2.11) [16], producing 1.73 Gb Illumina data. The *C. hemipterus* genome was assembled using the SPAdes (ver3.15.2) [17] genome assembler. After assembling the data, the length statistics of the genome assembly were assessed by QUAST (ver5.0.2) [18]. The quality and

genome completeness of the assembly was assessed using BUSCO (ver5.1.2) [13] against a set of highly conserved insect single-copy orthologs. The repetitive elements were identified and masked using the RepeatMasker (ver4.1.2) [19]. The number of protein-coding genes of the masked assembly was estimated using AUGUSTUS (ver3.4.0) [20]. The protein-coding genes were then annotated against the Eukaryotic Orthologous Groups (KOG) database using the WebMGA server [21].

Author Contributions: L.L. and A.H.A.M. designed the experiment; L.L. collected data, analyzed the data, and wrote the manuscript. A.H.A.M. read, corrected, and approved the manuscript. A.H.A.M. provided supervision and was responsible for project administration and funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Fundamental Research Grant (FRGS), 203/PBIOLOGI/6711681.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Universiti Sains Malaysia Research Ethics Committee (Human) JEPeM, Code:USM/JEPeM/19120868.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data have been submitted to the Sequence Read Archive (SRA) database with accession numbers SAMN18780127 and SAMN18780128 under BioProject, PRJNA640473 while the assembled data is available in the figshare repository (DOI: https://doi.org/10.6084/m9.figshare.18393767.v1).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Hwang, S.W.; Svoboda, T.J.; De Jong, I.J.; Kabasele, K.J.; Gogosis, E. Bed bug infestations in an urban environment. *Emerg. Infect. Dis.* **2005**, *11*, 533. [CrossRef] [PubMed]
2. Doggett, S.L.; Miller, D.M.; Lee, C.Y. *Advances in the Biology and Management of Modern Bed Bugs*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
3. Campbell, B.E.; Koehler, P.G.; Buss, L.J.; Baldwin, R.W. Recent documentation of the tropical bed bug (Hemiptera: Cimicidae) in Florida since the common bed bug resurgence. *Fla. Entomol.* **2016**, *99*, 549–551. [CrossRef]
4. Masini, P.; Zampetti, S.; Miñón Llera, G.; Biancolini, F.; Moretta, I.; Romani, R.; Tramontana, M.; Hansel, K.; Stingeni, L. Infestation by the tropical bedbug *Cimex hemipterus* (Hemiptera: Cimicidae): First report in Italy. *J. Eur. Acad. Dermatol. Venereol.* **2020**, *34*, 28–30. [CrossRef] [PubMed]
5. Chebbah, D.; Elissa, N.; Sereno, D.; Hamarsheh, O.; Marteau, A.; Jan, J.; Izri, A.; Akhoundi, M. Bed bugs (Hemiptera: Cimicidae) population diversity and first record of *Cimex hemipterus* in Paris. *Insects* **2021**, *12*, 578. [CrossRef] [PubMed]
6. Usinger, R.L. *Monograph of Cimicidae (Hemiptera–Heteroptera)*; Entomological Society of America: College Park, MA, USA, 1966.
7. Zulaikha, Z.; Hassan, A.A. A survey on the infestation levels of tropical bed bugs in Peninsular Malaysia: Current updates and status on resurgence of *Cimex hemipterus* (Hemiptera: Cimicidae). *Asian Pac. J. Trop. Dis.* **2016**, *6*, 40–45. [CrossRef]
8. Doggett, S.L.; Dwyer, D.E.; Peñas, P.F.; Russell, R.C. Bed bugs: Clinical relevance and control options. *Clin. Microb. Rev.* **2012**, *25*, 164–192. [CrossRef] [PubMed]
9. Pritchard, M.J.; Hwang, S.W. Severe anemia from bedbugs. *CMAJ* **2009**, *181*, 287–288. [CrossRef] [PubMed]
10. Rahim, A.H.A.R.; Zahran, Z.; Majid, A.H.A. Human skin reactions towards bites of tropical bed bug, *Cimex hemipterus* F. (Hemiptera: Cimicidae): A preliminary case study. *Asian Pac. J. Trop. Dis.* **2016**, *6*, 366–371. [CrossRef]
11. Effiom, O.E.; Akuwudike, C.C. Investigation of *Cimex* (Bedbugs) infestation in human settlements in Bwari Area Council, Abuja FCT, Nigeria. *World J. Innov. Res.* **2021**, *10*, 13–20. [CrossRef]
12. Deku, G.; Combey, R.; Doggett, S.L.; Mensah, B.A. Assessment of tropical bed bug (Hemiptera: Cimicidae), infestations in Cape Coast, Ghana: Household control practices and efficacy of commercial insecticides and long-lasting insecticidal nets against field bed bugs. *J. Med. Entomol.* **2021**, *58*, 1788–1797. [CrossRef] [PubMed]
13. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef] [PubMed]
14. Andrews, S.; Krueger, F.; Seconds-Pichon, A.; Biggins, F.; Wingett, S. FastQC: A quality control tool for high throughput sequence data. *Babraham Bioinform.* **2010**. Available online: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed on 16 May 2022).

15. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114–2120. [CrossRef] [PubMed]
16. Magoč, T.; Salzberg, S.L. FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **2011**, *27*, 2957–2963. [CrossRef] [PubMed]
17. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Pevzner, P.A.; et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **2012**, *19*, 455–477. [CrossRef] [PubMed]
18. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef] [PubMed]
19. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinform.* **2004**, *5*, 4–10. [CrossRef] [PubMed]
20. Stanke, M.; Morgenstern, B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **2005**, *33*, 465–467. [CrossRef] [PubMed]
21. Wu, S.; Zhu, Z.; Fu, L.; Niu, B.; Li, W. WebMGA: A customizable web server for fast metagenomic sequence analysis. *BMC Genom.* **2011**, *12*, 444. [CrossRef] [PubMed]