data

MDPI

Data Descriptor

Isolation of Microsatellite Markers from De Novo Whole Genome Sequences of *Coptotermes gestroi* (Wasmann) (Blattodea: Rhinotermitidae)

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

[1] Household & Structural Urban Entomology Laboratory, Vector Control Research Unit, School of Biological Sciences, Universiti Sains Malaysia, Minden 11800, Penang, Malaysia; yang940715@gmail.com (L.Y.L.); shawn@frim.gov.my (S.C.)
[2] Genetics Laboratory, Forest Research Institute Malaysia (FRIM), Biotechnology Division, Kepong 52109, Selangor, Malaysia
* Correspondence: abdhafiz@usm.my

Abstract: *Coptotermes gestroi* (Wasmann) (Blattodea: Rhinotermitidae) is a subterranean termite species from Southeast Asia which has been unintentionally introduced to many parts of the world through commerce and modern transportation. Known for causing extensive damage to timber used in the built environment, the termite also has a habit of nesting in carton nests in wood and wooden structures in buildings. As so little is known of its breeding system, colony, and genetic structure, we initiated work to sequence its genome with an Illumina HiSeq™ 2000 sequencer. In this publication, we announce our paired-end sequencing data and report the isolation of 119,190 microsatellite markers from our DNA assembly. The microsatellite marker reported in this publication can be used to elucidate the mating system and genetic structure of this highly invasive termite species. Additionally, in this announcement the study authors make the Bio Project sequence accession number SRR13105492 accessible from the Sequence Read Archive database.

Dataset: https://www.ncbi.nlm.nih.gov/sra/SRR13105492.

Keywords: *Coptotermes gestroi*; microsatellites; termites; genomic DNA; whole-genome sequencing; Illumina HiSeq™ 2000

Citation: Lim, L.Y.; Cheng, S.; Ab Majid, A.H. Isolation of Microsatellite Markers from De Novo Whole Genome Sequences of *Coptotermes gestroi* (Wasmann) (Blattodea: Rhinotermitidae). *Data* **2021**, *6*, 40. https://doi.org/10.3390/data6040040

Received: 22 March 2021
Accepted: 7 April 2021
Published: 10 April 2021

1. Summary

Coptotermes gestroi (Wasmann) (Blattodea: Rhinotermitidae) is a highly invasive termite species and a major pest of both the timber and wood used in buildings. These data aim to contribute to genomic DNA data on termites, particularly from genus *Coptotermes*. The data source was located at Penang island, Malaysia, with GPS Coordinates 5°21′13.716″ N, 100°18′5.112″ E. The DNA data were acquired through whole-genome DNA sequencing on the Illumina HiSeq™ 2000 sequencing system. The DNA data were used to isolate microsatellite DNA markers for population genetic analysis of *C. gestroi*. The markers may be cross amplified in other *Coptotermes* spp. The full list of the 119,190 microsatellite primers designed from our dataset can be found in Supplemental Information S1.

2. Data Description

2.1. Raw Sequence Reads

Our dataset contains paired-end DNA sequence reads (1_R1.fq and 1_R2.fq) from *Coptotermes gestroi* generated by an Illumina HiSeq 2000 sequencer (150 nucleotide/base pair). The dataset was deposited in NCBI's SRA database, with the accession number SRR13105492, under BioProject number PRJNA679986.

File names/numbers "1_R1.fq" and "1_R2.fq" contain forward sequence reads (150 bp in length) and reverse sequence reads (between 150 bp and 300 bp in length), respectively. We obtained a total of 34,444,724 sequence reads at an error rate of 0.0295% (see Table 1). A large number of our sequences, that is, 90.59% of the 5,166,708,600 bases that were sequenced, had a Q30 Phred score with the dataset having an overall Guanine-Cytosine content (GC-content) of 41.33% (Table 1).

Table 1. Paired-end sequencing data statistics for *C. gestroi*.

Total # of Reads	Total # of Bases	Total # of Reads with N's (%)	Error%	Q20%	Q30%	GC%
34,444,724	5,166,708,600	20,354 (0.06)	0.0295	95.96	90.59	41.33

2.2. Microsatellite Markers

In total, we detected 3.57×10^6 microsatellite loci with the following motif types: 2.33×10^6 mononucleotide; 3.60×10^5 dinucleotide; 2.38×10^5 trinucleotide; 5.25×10^5 tetranu-cleotide; 1.12×10^5 pentanucleotide; and finally, 1.48×10^3 hexanucleotide. From these, we were able to design a total of 119,190 primers/markers for 71,949 mononucleotide, 20,373 dinucleotide, 11,942 trinucleotide, 13,790 tetranucleotide, 1070 pentanucleotide, and finally, 66 hexanucleotide microsatellite loci. A selection of microsatellite markers is shown in Table 2. However, further validation through polymorphism analysis is required to characterize the primers for genetic population studies. Although work to characterize a subset of these markers is currently on-going at Universiti Sains Malaysia, Penang, we encourage others to also undertake/publish microsatellite marker validation experiments using the markers we have isolated so we can better understand the biology of *Coptotermes* spp. The full list of the 119,190 microsatellite primers designed from our dataset can be found in Supplemental Information S1.

Table 2. Selection of microsatellite markers designed for *C. gestroi*. See Supplemental Information S1 for full list of markers.

ID No.	Left Primer	Left T_m (°C)	Left GC (%)	Right Primer	Right T_m (°C)	Right GC (%)	Repeat Motif
TMDi01	TAATGGGACCGAATGTTGGC	59.0	50.0	CGCTGGAGTGTTGTCGTTAC	60.0	55.0	$(AC)_{46}$
TMDi02	CAACGAACCAACCACACACC	60.6	55.0	TCATTCTCCGTGAGTTTGCG	59.3	50.0	$(AC)_{44}$
TMDi03	AGCTCTCACCTTTATCGACCC	60.0	52.4	TGGTTGTGTCTTTGTGGCTC	59.4	50.0	$(AC)_{43}$
TMDi04	TGTGACGGACGAAATGACAG	60.0	50.0	ATCGGTGCCTAGTGAGAGTG	59.7	55.0	$(AC)_{42}$
TMDi05	GCCACCTCATAAGCTGTTCG	59.7	55.0	TTTCTGTGGCTGTGGTTTGG	59.6	50.0	$(AC)_{43}$
TMTr01	GGAGGGAGGCAGGGAATTC	60.2	63.2	CCACCACGAACTATAACGCC	59.5	55.0	$(AAG)_{23}$
TMTr02	ACGAAGGTGTAGCTGGAGTG	60.2	55.0	GCTCACCTCCATTAACTGCC	59.4	55.0	$(AAG)_{25}$
TMTr03	ACGCCACACTTTAAACCCTC	59.2	50.0	TTCTCGCTTCCACCTCACAC	60.7	55.0	$(AAC)_{26}$
TMTr04	AGGCTCGATGTAGAACAGGG	59.6	55.0	CAGTCTTGTTTGCGTGAGGG	60.2	55.0	$(AAG)_{22}$
TMTr05	GACCCGCCTGAGTGTAAATG	59.4	55.0	AGGCCTGTGACAAGAGACTC	59.8	55.0	$(AAT)_{23}$
TMTe01	TGGCATAGTGTCGGTGGTAG	59.9	55.0	TCAGTATCATTGCGCCAGTG	59.9	50.0	$(AATG)_{23}$
TMTe02	CGTCCAGTGAGAGCCAAATTG	60.3	52.4	GCGACCATTAGCTACACCAC	59.5	55.0	$(ACAG)_{23}$
TMTe03	TCGTATCTGTCCGTCCATCTG	59.8	52.4	GTTACGCCGATACTTGCTGG	59.8	55.0	$(AGGC)_{22}$
TMTe04	TAGCCGAGTTGTTTCATGCG	59.7	50.0	GTCCCTATGTCTGTCCGTCC	60.0	60.0	$(ACAG)_{20}$
TMTe05	ATCTGTCTATTGGGAGCCGC	60.4	55.0	GGAGCCTTCGTTTAATCGCG	60.4	55.0	$(ACAG)_{20}$
TMPe01	TTGGAAGAGAAATGCGCGAC	59.9	50.0	GCGAATGATACTCCAGAAGCG	59.9	52.4	$(AACAC)_{18}$
TMPe02	GCGATTTGACAGCCACCTC	59.9	57.9	GGCATTCTGGGTTTGTGCC	60.5	57.9	$(AATAG)_{16}$
TMPe03	ACAGGCGGAAGAAGTGATGG	60.8	55.0	GTGTAGTTCCGTGTCGCTTC	59.7	55.0	$(AAGGG)_{12}$
TMPe04	AACCCGGACAGTTTAGCCC	60.4	57.9	TCGACCCACTGTTTGTTTGC	60.0	50.0	$(AACAG)_{11}$
TMPe05	AGGGCATGTGACAGGTATGG	60.2	55.0	AACACTGCAAGCTCCTCTCC	60.7	55.0	$(AGAGG)_{13}$

2.3. Value of the Data

We have made available de novo next-generation sequencing data containing raw, paired-end sequencing reads of the *C. gestroi* genome from which we were able to isolate 119,190 microsatellite markers. The markers may be cross amplified in other *Coptotermes* spp. These data can be used later for studying the genetic structure of *C. gestroi* populations. *Coptotermes gestroi* is a highly invasive termite species and a significant pest of both timber and wood used in buildings.

3. Methods

3.1. Termite Sampling and Laboratory Protocols

Specimens of the subterranean termite, *Coptotermes gestroi*, were collected from an underground monitoring station filled with *Pinus caribaea* following the protocol of Ab Majid and Ahmad [1]. Total genomic DNA (gDNA) was, however, extracted from a single *C. gestroi* soldier (instead of a termite worker), using only its head capsule (tissue) to minimize contamination from endosymbionts [2]. The termite head capsule was dissected away from the thorax and abdomen and rinsed with 70% ethanol and sterilized water to remove any possible contaminants found externally on the head capsule. We used HiYield Plus™ Genomic DNA Mini Kit (Blood/Tissue/Cultured Cells) (Real Biotech Corp., Taipei, Taiwan) to extract gDNA according to the manufacturers' instructions for cultured cells. Briefly, the head capsule was crushed with a microtube pestle after which 200 µL of GB Buffer was added to the solution and vortexed. After incubation at 60 °C for 1 h, ethanol was added to the lysate and the solution was transferred into a spin column for DNA binding. After applying a wash buffer to the spin column, 100 µL of pre-heated TE buffer was added to the center of the column matrix and left to stand for several minutes after which it was centrifuged at 14,000–16,000× *g* for 30 s to elute the purified DNA. The DNA quantity and quality were evaluated by spectrophotometry on a NanoDrop 2000c Spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA).

3.2. Library Preparation and Sequencing

Samples were submitted to Apical Scientific Sdn. Bhd. (Selangor, Malaysia) for DNA library preparation and sequencing. The DNA library was prepared with NEB Next® DNA Library Prep Kit (New England Biolabs Inc., Ipswich, MA, USA) by shearing/digesting the DNA into 350 bp fragments and end-repairing the fragments with a dA-tail. The DNA fragments were then ligated with NEBNext® Adapter(s) and amplified via polymerase chain reaction using P5- and P7-indexed primers. The sequences were then purified with AMPure XP system (Beckman Coulter, Indianapolis, IN, USA). Size distribution and quantity validation were performed on an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA) and real-time PCR, respectively. Finally, the qualified DNA library was sequenced on an Illumina HiSeq™ 2000 sequencing system.

3.3. Microsatellite Markers Design

We first isolated microsatellite from our DNA assembly with Msatcommander v1.0.8 [3], following which we designed forward and reverse primers in Primer3Plus [4]. The minimum requirement of perfect repeats in isolating microsatellite for each motif types is set as 8 for mononucleotide, 8 for dinucleotide, 8 for trinucleotide, 6 for tetranucleotide, 6 for pentanucleotide, and 6 for hexanucleotide. The primers designed based on isolated microsatellites are restricted between 18 to 22 bp primer sizes, a melting temperature of 58 to 62 degrees Celsius and 30% to 70% GC content. The selection of designed microsatellite markers in Table 2 was performed based on the microsatellite length, repeat motif and penalty score as shown in Supplemental Information S1.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/data6040040/s1, The full list of the 119,190 microsatellite primers designed from our dataset can be found in Supplemental Information S1.

Author Contributions: Conceptualization: A.H.A.M.; methodology, A.H.A.M., S.C. and L.Y.L.; validation, A.H.A.M., S.C. and L.Y.L.; formal analysis, L.Y.L.; investigation, A.H.A.M., S.C. and L.Y.L.; resources, A.H.A.M.; data curation, A.H.A.M., S.C. and L.Y.L.; writing—review and editing A.H.A.M., S.C. and L.Y.L.; supervision, A.H.A.M.; project administration, A.H.A.M.; funding acquisition, A.H.A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a Universiti Sains Malaysia, Research University Grant Rui (Grant No: 1001/PBIOLOGI/8011104).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: BioProject: PRJNA679986 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA679986), accessed on 25 March 2021; BioSample: SAMN16868959 (https://www.ncbi.nlm.nih.gov/biosample/SAMN16868959), accessed on 25 March 2021; NCBI SRA: SRR13105492 (https://www.ncbi.nlm.nih.gov/sra/SRR13105492), accessed on 25 March 2021.

Acknowledgments: We would like to thank Universiti Sains Malaysia, Research University Grant Rui (Grant No: 1001/PBIOLOGI/8011104) for funding this project and Forest Research Institute Malaysia (FRIM) for technical support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ab Majid, A.H.; Ahmad, A.H. Define colony number of subterranean termites *Coptotermes gestroi* (Isoptera: Rhinotermitidae) in selected infested structures. *Sains Malays.* **2015**, *44*, 211–216. [CrossRef]
2. Moreau, C.S. A practical guide to DNA extraction, PCR, and gene-based DNA sequencing in insects. *Halteres* **2014**, *5*, 32–42.
3. Faircloth, B. Msatcommander: Detection of microsatellite repeats arrays and automated, locus-specific primer design. *Mol. Ecol. Resour.* **2008**, *8*, 92–94. [CrossRef] [PubMed]
4. Untergasser, A.; Cutcutache, I.; Koressaar, T.; Ye, J.; Faircloth, B.; Remm, M.; Rozen, S. Primer3—New capabilities and interfaces. *Nucleic Acids Res.* **2012**, *40*, e115. [CrossRef] [PubMed]