Indian Journal of Entomology Online published Ref. No. e24371 DoI. No.: 10.55446/IJE.2024.2371

DE NOVO GENOME ASSEMBLY OF THE MAIN MALARIA VECTOR *ANOPHELES BAIMAII*

Sedthapong Laojun[1], Pongmada Damapong[1], Peerada Damapong[1], Tanawat Chaiphongpachara[1]*

[1]Department of Public Health and Health Promotion, College of Allied Health Sciences, Suan Sunandha Rajabhat University, Samut Songkhram, Thailand 750008
*Email: tanawat.ch@ssru.ac.th (corresponding author): ORCID ID 0000-0001-9585-4241

ABSTRACT

***Anopheles baimaii*, a primary vector of human malaria in Southeast Asia's forested regions, including Thailand, plays a pivotal role in pathogen transmission. The lack of a reference genome for *An. baimaii* limits our comprehensive understanding of its biology. This study presents the first genome assembly for *An. baimaii*, consisting of 1,098,224 contigs and exhibiting a GC content of 46.4%. *k-mer* analysis estimates the genome size at 443 megabases (Mb), a finding corroborated by BLAST results that align the lengths of the top five contigs with those of other *Anopheles* spp., as verified by comparison with genomes in the NCBI NT database. Validation of the assembly through Illumina read mapping achieved a 94.78% mapping rate but revealed a low average mapping depth of 47.44. Furthermore, BUSCO analysis indicated a low degree of completeness, with only 23.92% of BUSCOs completed. Despite these challenges, this draft genome assembly provides a crucial reference for *An. baimaii* and lays the groundwork for future molecular biology research and malaria control strategies.**

Key words: *Anopheles baimaii*, Thailand, genome assembly, malaria vector, malaria, mosquito-borne disease, *k-mer*, blast, mapping depth, BUSCO analysis

Anopheles baimaii (Diptera: Culicidae), a member of the *Dirus* complex and formerly known as *An. dirus* species D, serves as the primary vector of malaria (O'Loughlin et al., 2008). This species frequently occurs in forested and forest-fringe zones in Southeast Asia, spanning from the northeastern part of India, through the hills of Bangladesh, across Myanmar, and into the northwestern and southern parts of Thailand (Obsomer et al., 2007; O'Loughlin et al., 2008). This mosquito species is an important target for vector control to reduce the risk of malaria in many countries, especially in the Myanmar–Thailand border areas (Parker et al., 2015). This *Anopheles* species exhibits a preference for humans as a blood source (anthropophilic behaviour) and tends to rest outdoors (exophagic behaviour) (Suwonkerd et al., 2013). Its breeding sites are frequently small, shallow, and usually temporary, consisting of mostly shaded bodies of fresh, stagnant, or slowly flowing water (Hii and Rueda, 2013). These include rainwater pools, puddles, animal footprints (e.g., elephant or buffalo footprints), and artificial containers in forested mountain and foothill areas. Consequently, *An. baimaii* is more prevalent in the rainy season compared to other seasons, as numerous natural breeding grounds are available. Additionally, this species is sometimes found in dense mono-agricultural environments, particularly in rubber, fruit, and cassava plantations in certain areas (Suwonkerd et al., 2013).

Morphological identification of members within the *Dirus* complex is challenging due to their very similar characteristics (Walton et al., 1999; O'Loughlin et al., 2008). The allele-specific polymerase chain reaction (AS-PCR) method, based on the internal transcribed spacer 2 (ITS2) sequence, is commonly used for this purpose (Walton et al., 1999). Although DNA barcoding, based on the cytochrome *c* oxidase subunit 1 (*COI*), is another method for identifying mosquitoes with similar morphology, it has proven ineffective in distinguishing between closely related species, including *An. dirus* and *An. baimaii* (Chaiphongpachara et al., 2022a; Chaiphongpachara et al., 2022b). Currently, the PCR technique for classifying *Anopheles* members in this complex has been improved to be more efficient and is known as the *Dirus* Complex Species Identification PCR (DiCSIP) (Saeung et al., 2023). Additionally, a recent report suggests that geometric morphometrics, which analyze differences in wing shape, may offer potential in differentiating between *An. dirus* and *An. baimaii* (Chaiphongpachara et al., 2022b).

For other molecular biology applications, investigation of insecticide resistance at the molecular

level is widely used in several *Anopheles* spp., including *An. gambiae* (Mwagira-Maina et al., 2021), *An. coluzzii* (Lucas et al., 2023), *An. funestus* (Weedall et al., 2020), and *An. baimaii* (Chaiphongpachara et al., 2022c). In addition, examining genetic variation in *Anopheles* mosquitoes is also crucial for understanding their population structure, aiding in mosquito control and tracking traits such as vector competence and insecticide resistance (Kaddumukasa et al., 2020; Sumitha et al., 2023). A previous study on genetic variation revealed that *An. baimaii* populations in Thailand and Myanmar originated from a dispersal from Northeast India approximately 135,000 to 173,000 years ago (O'Loughlin et al., 2008). This migration led to a bottleneck effect in mitochondrial DNA upon departure from Northeast India, resulting in distinct genetic differentiation of *An. baimaii* populations in Bangladesh, Myanmar, and Thailand (Sarma et al., 2012). Therefore, several past studies have proven that investigations at the molecular biology level are extremely important in monitoring and controlling malaria vector mosquitoes, such as the primary malaria vector, *An. baimaii*.

Currently, the next-generation sequencing (NGS), often referred to as high-throughput sequencing, is an emerging technique that not only enables extensive genome sequencing, including whole genome sequencing and de novo sequencing, but also supports transcriptomics, epigenomics, metagenomics, and other omics studies (Satam et al., 2023). This technology has significantly advanced the investigation of genetic variation, gene expression, and epigenetic changes, leading to interesting discoveries (Satam et al., 2023). De novo sequencing is sequencing an organism's genome for the first time without using a reference genome for alignment (Lischer and Shimizu, 2017). In this process, sequence reads are assembled into contigs. The coverage quality of de novo sequence data depends on the size and continuity of these contigs, specifically the frequency of gaps in the data (Dida and Yi, 2021). In our study, we performed the first genome assembly for the primary malaria vector, *An. baimaii*. This genome assembly will serve as a valuable reference for *An. baimaii*, forming the basis for future molecular biology studies involving DNA sequences and being crucial for advancing malaria control efforts.

MATERIALS AND METHODS

In this study, mosquitoes were collected from three sites in Ranong Province, Southern Thailand: site 1 (10°45′48.3″ N, 98°53′31.4″ E), site 2 (9°57′20.1″ N, 98°42′05.3″ E), and site 3 (9°22′06.8″ N, 98°27′52.1″ E). This collection was based on a report that previously identified the presence of *An. baimaii* at these locations (Chaiphongpachara et al., 2022c). Three Centers for Disease Control and Prevention (CDC) light traps (John W. Hock Co., Gainesville, FL, USA) with dry ice (solid carbon dioxide) were used to collect *Anopheles* mosquitoes between February and June 2022. All mosquitoes were preserved in 1.5 ml Eppendorf tubes (10 samples/ tube) with 1 ml of RNAlater (Thermo Fisher Scientific, Waltham, MA, United States) and sent to the College of Allied Health Sciences Laboratory at Suan Sunandha Rajabhat University, Samut Songkhram Campus, for species identification.

The *An. dirus* complex was identified morphologically using the standard taxonomic key (Rattanarithikul et al., 2006) under a Nikon SMZ 800 N stereomicroscope (Nikon Corp., Tokyo, Japan). For species identification within the complex, two legs were removed from all samples identified as *An. dirus* complex for DNA extraction. Genomic DNA was then extracted using the FavorPrep™ Mini Kit (Favorgen Biotech, Ping-Tung, Taiwan), following the manufacturer's instructions. The remaining parts of each mosquito sample were subsequently stored individually in 1.5 ml Eppendorf tubes with 50 µl of RNAlater and kept at -20 °C. A multiplex PCR assay, based on ITS2 and using primers specific to *An. dirus* s.s., *An. baimaii*, *An. cracens, An. nemophilous*, and *An. scanloni*, was performed according to the procedure described by Walton et al. (1999). Following this, 30 samples identified as *An. baimaii* (10 samples/ tube) were sent to Macrogen Inc., Seoul, South Korea, for genome sequencing. This research protocol was reviewed, and approved by the Suan Sunandha Rajabhat University Institutional Animal Care and Use Committee (Ethics Approval Number: IACUC 64-010/2021).

After extracting DNA from the sample, a sample library was produced through random fragmentation of the DNA, followed by ligation of adapters to both the 5' and 3' ends. Subsequent to this, cluster generation was performed. The sample library was loaded onto a flow cell, where DNA fragments were embedded on a surface coated with oligonucleotides complementary to the library adapters. Each fragment undergoes amplification into separate, clonal clusters using either bridge amplification or exclusion amplification. Once the cluster creation process was completed, the templates were ready for sequencing. The sequencing

method, based on Illumina's Sequencing by Synthesis (SBS), employs a proprietary reversible terminator-based approach to detect individual bases as they are incorporated into DNA template strands. DNA polymerase, primers, and fluorescently labeled terminators were given to the flow cell. The primer then attaches to the DNA being sequenced, and DNA polymerase binds to this primer, incorporating the initial fluorescently labeled terminator into the newly formed DNA strand. After sequencing, each base that has been sequenced is converted into raw data (FASTQ file) for further analysis.

The raw sequence data reads were subjected to quality control inspection using FASTQC (http://www.bioinformatics. babraham.ac.uk/projects/fastqc/), and Trimmomatic software version 0.36 (Bolger et al., 2014) was used to remove incomplete adapters. *k-mer* analysis, which refers to the examination of subsequences of length K within a specific sequence, allows for the estimation of approximate genome size and heterozygosity rate in silico. Therefore, the genome properties of *An. baimaii*, including genome size, average coverage, and heterozygosity, were estimated using *k-mer* analysis conducted with GenomeScope software (Vurture et al., 2017). De novo assembly involves breaking down the entire genome into short DNA sequences, then sequencing these fragments, and finally assembling the sequencing data to reconstruct the genome sequence. In this study, de novo assembly of the filtered reads was carried out using SOAPdenovo2 software version 2.04, which is based on a novel short-read assembly method (Luo et al., 2012). The best *k-mer* was selected based on assembled results, including the number of contigs, the total length of all contigs, contig sum, N50, and others. The term 'contig' refers to the contiguous DNA segments that represent an organism's chromosomes, which are the final output of de novo assembly using sequencing reads (Chakraborty et al., 2016). The N50 metric (or Nx), denotes the length of a contig at which 50% (or x%) of the total genome length is covered when contigs are arranged in descending order of length and is a key measure for assessing the completeness of the assembly (Alhakami et al., 2017).

Upon completing the assembly, the resulting data were utilized to validate the completeness of the draft genome. The genome assembly of *An. baimaii* underwent validation through two approaches: self-mapping and Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis (Simão et al., 2015). In the self-mapping process, filtered reads were aligned against the assembled genome to estimate their insert size and assess the proportion of reads incorporated into the assembly, with the raw data reads being mapped back to the assembly results. Simultaneously, BUSCO analysis was conducted to assess the genome assembly's completeness, relying on evolutionarily informed expectations of gene content derived from near-universal single-copy orthologs. Matches are categorized as "complete" if their lengths correspond to the anticipated lengths according to the BUSCO profile. Matches found multiple times are classified as "duplicated." Those only partially recovered are labeled "fragmented," and BUSCO groups without matches that fulfill the orthology tests are considered "missing." The genome assembly has been deposited in the BioProject under PRJNA1070618: *Anopheles baimaii* Genome Sequencing and Assembly (TaxID: 48725), accessible through the GenBank database.

RESULTS AND DISCUSSION

Before proceeding with genome assembly, *k-mer* distribution and analysis were conducted to estimate the genome size of the sample. This study presents results using a *k-mer* size of 21 (21-mer). The *k-mer* distribution graph revealed a prominent peak at approximately 19x coverage and a smaller peak around 78x, as shown in Fig. 1. Due to overdispersion in the real data, the top of the peak does not intersect with

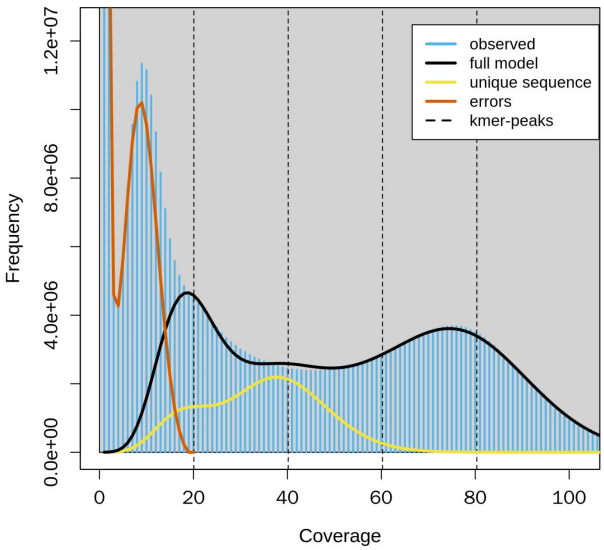

Fig. 1. *k-mer* profile (k = 21) spectrum analysis used to estimate the genome size in *An. baimaii*, generated from sequence data using GenomeScope. In this *k-mer* graph, the coverage and frequency of *k-mers* are plotted. The sharp peak on the left side represents random sequencing errors, while the peak on the right indicates appropriate data

the *k-mer* peaks line. The first peak corresponds to heterozygous single-copy *k-mers*, while the second peak is indicative of homozygous single-copy *k-mers*, both of which are instrumental in estimating genome size. Based on the total *k-mer* count and the volume of these peaks, the genome size of *An. baimaii* was estimated to be 443,108,634 bases. This includes a genome repeat length of 382,707, 525 bases and a mean *k-mer* coverage of 40.18 for heterozygous bases. The overall heterozygosity rate was determined to be 0.942. The estimated genome size of *An. baimaii*, as determined by *k-mer* analysis, is 443 megabases (Mb), where one Mb equals one million base pairs of DNA. This analysis reveals that the *An. baimaii* genome is larger than those of several other *Anopheles* species: it exceeds the 221 Mb genome of *An. stephensi* (Jiang et al., 2014), the 280 Mb genome of *An. gambiae* (Holt et al., 2002), the 395 Mb genome of *An. cracens*, the 137 Mb genome of *An. darlingi*, and the 375.8 Mb genome of *An. sinensis* (Lau et al., 2016). However, it is comparable in size to the 499 Mb genome of *An. maculatus*, suggesting a similarity in high repeat content as a likely reason for the large genome size (Lau et al., 2016). Furthermore, in comparison with other mosquito species, the *An. baimaii* genome is smaller than the 579 Mb genome of *Cx. quinquefasciatus* (Arensburger et al., 2010) and significantly smaller than the 1376 Mb genome of *Ae. aegypti* (Nene et al., 2007).

De novo assembly was conducted with various *k-mers* using SOAPdenovo2. The optimal *k-mer* was selected based on assembly outcomes, taking into account factors such as the number of contigs, the total number of bases in the contigs (contig sum), and the N50 metric, which indicates that half of all bases are in contigs of this size or longer. The resulting draft genome assembly is comprised of 1,098,224 contigs, with a GC content of 46.4%. An assembly summary is provided in Table 1, and the base contents of the contigs are elaborated in Table 2. The lengths of the top 10 longest contigs of *An. baimaii* are depicted in Fig. 2. The *An. baimaii* genome exhibits a GC content of 46.4%, consistent with findings from previous studies on *Anopheles* genomes, which reported a GC

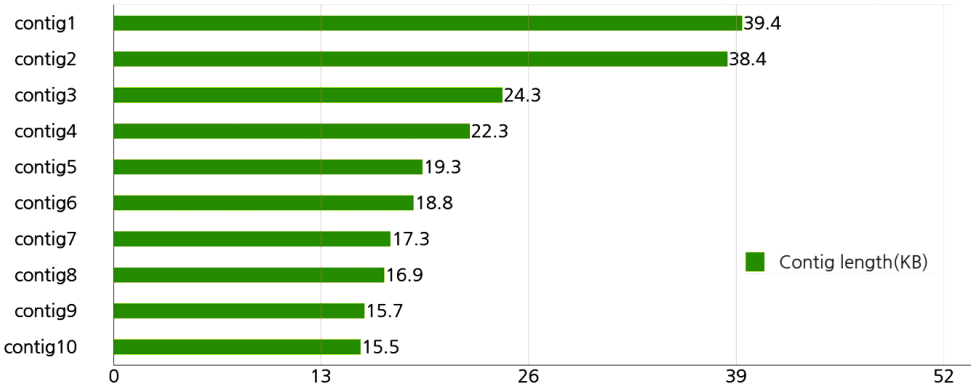

Fig. 2. Length of the top 10 longest contigs in the *An. baimaii* assembly. The unit of measurement in the figure is kilobases (Kb), with one Kb equaling one thousand base pairs of DNA

Table 1. Assembly summary of contigs in the *An. baimaii* assembly

No. of contigs	Contigs sum	N50	Longest contig	Shortest contig	Average length
1,098,224	397,570,932	453	39,347	124	362

No. of contigs = the number of contigs identified; Contigs sum = the total number of bases in the contigs; N50 = a N50 means that half of all bases reside in contig of this size or longer; Average length = Average contig size.

Table 2. Base contents of contigs in the *An. baimaii* assembly

Num of A	Num of T	Num of G	Num of C	Num of N	GC contents
105,916,164	103,559,789	90,210,435	91,699,465	6,185,079	46.4%

Num of A = The total number of adenine (A); Num of T = The total number of thymine (T); Num of G = The total number of guanine (G); Num of C = The total number of cytosine (C); Num of N =The total number of ambiguous base; GC contents = The percentage of guanine-cytosine base pairs.

Table 3. Overall mapping result of the *An. baimaii* assembly

Library name	Total reads	Mapped reads	Coverage (%)	Depth	Ins.size (Std.)
An. baimaii	144,646,582	37,036,727 (94.74%)	94.78	47.44	431.86 (112.14)

Table 4. BLAST result of top 5 contigs of the *An. baimaii* assembly

Contig	Contig Length	Subject Description	Subject Length	E-Value
Contig 1	39,347	M_040385588.1 PREDICTED: *Anopheles coluzzii* uncharacterized (LOC120961659), mRNA	14,452	1e-124
Contig 2	38,425	XM_041923284.1 PREDICTED: *Anopheles merus* smoothelin-like (LOC121597475), transcript variant X4, mRNA	12,081	2e-141
Contig 3	24,321	XM_040379819.1 PREDICTED: *Anopheles coluzzii* protein-tyrosine sulfotransferase (LOC120957544), transcript variant X3, mRNA	4,724	1e-73
Contig 4	22,278	CP032300.1 *Anopheles stephensi* strain Indian chromosome 2L	25,562,539	7e-80
Contig 5	19,323	CP032301.1 *Anopheles stephensi* strain Indian chromosome 3R	43,030,556	2e-55

Note: Subject length (bp) = length of the sequence matched by BLASTN; E-value = The expectations that could be matched by chance. The lower, the more significant it is.

content of 42.6% in *An. sinensis* (Zhou et al., 2014) and 44.91% in *An. stephensi* (Chakraborty et al., 2020). Research suggests that GC content correlates with intron numbers, potentially influencing the genetic diversity rate (Zhou et al., 2014). The draft genome of *An. baimaii* comprises 1,098,224 contigs, with many being smaller-sized contigs (N50 = 453 bp). The presence of a large number of small contigs could indicate suboptimal genome quality. Nevertheless, this genomic data is highly valuable, classified as 'genetic data' (Ellis et al., 2021). This assertion is supported by BLAST results indicating that the lengths of the top five contigs correspond with those of *Anopheles* species, based on comparisons with genomes in the NCBI NT database. Moreover, it is common that despite the presence of numerous small contigs, the majority of the assembled sequence might be comprised within a few larger contigs (Wang et al., 2002).

To estimate the insert size of the raw data and the proportion of reads used in the assembly, the raw data reads were mapped to the assembly results. Following this mapping, several metrics were calculated: the percentage of mapped sites (% coverage), the average mapping depth (depth), and the length between adapters along with the standard deviation of the predicted length (ins. size). These results are presented in Table 3. The completeness of the genome assembly was assessed using BUSCO analysis, based on evolutionary expectations of gene content from nearly universal single-copy orthologs. The BUSCO analysis results indicated a relatively low degree of completeness, with 23.92% of the genes identified as complete BUSCOs, 45.88% as fragmented BUSCOs, and 30.20% as missing BUSCOs (Fig. 3). After assembling the complete genome, a BLAST analysis was conducted to determine the species similarity for each scaffold. The best hit and top 5 hits were identified using the NCBI NT database. The results, categorized at the genus level based on the best hit for the entire contig, are displayed in Fig. 4. Subsequently, the best hit BLAST results for up to 5 contigs, arranged by length, are presented in Table 4.

The self-mapping results indicated that 94.78% of the sites were successfully mapped, albeit with a low average mapping depth of 47.44. Generally, a higher

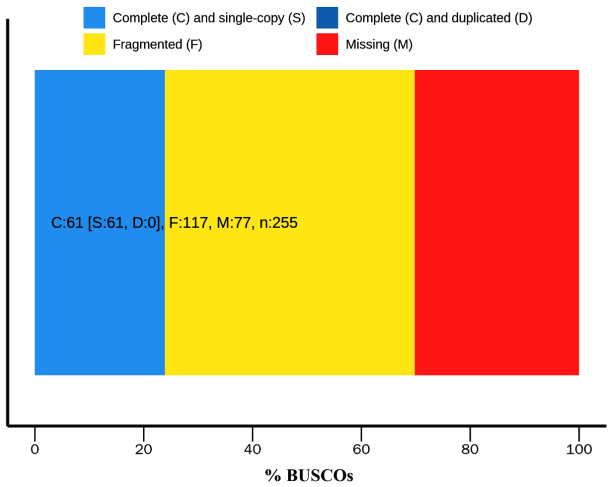

Fig. 3. BUSCO bar plot of in the *An. baimaii* assembly

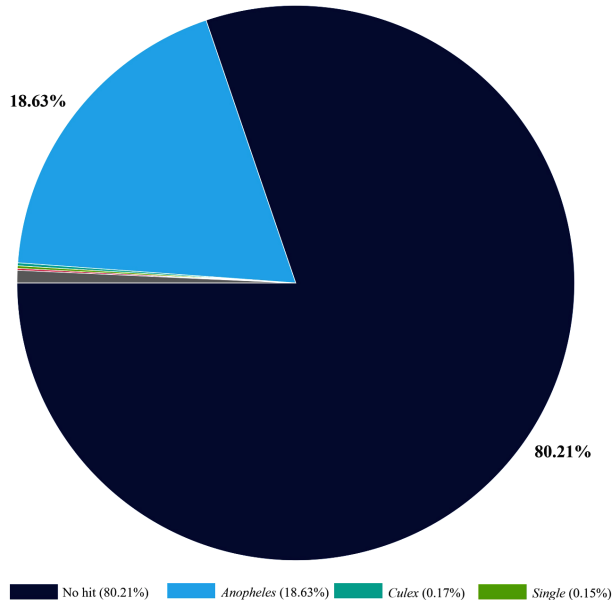

Fig. 4. Genus level summary of BLAS analysis

sequencing depth-indicating more frequent reading of the genome-yields more accurate and reliable information. The observed low mapping depth can likely be attributed to the presence of a large number of small contigs in the assembly. In terms of assessing the assembly's completeness, the BUSCO analysis revealed a relatively low degree of completeness, with only 23.92% of BUSCOs classified as complete. Typically, there is a correlation between the BUSCO score and the N50 value; a higher figure in one metric usually suggests a higher figure in the other. However, the analysis identified a significant rate of fragmented BUSCOs, suggesting potential issues in sequencing or assembly processes, possibly due to the abundance of small contigs (Manni et al., 2021). Therefore, for those intending to use this assembly as a reference, it is advisable to filter out the smaller contigs before further application. This study marks the first successful generation of a draft whole genome for the malaria vector, *An. baimaii*, through de novo assembly. Although the assembled genome of *An. baimaii* revealed a large number of small contigs, these features could either reflect the inherent genetic complexity of this mosquito species or arise from limitations in the assembly process. Nonetheless, this genetic data is invaluable, offering a wealth of genetic insights into this mosquito species.

ACKNOWLEDGEMENTS

This work was supported by the Suan Sunandha, Rajabhat University, Thailand.

AUTHOR CONTRIBUTION STATEMENT

S L and T W were responsible for conceptualization, investigation, methodology, data curation, writing-review and editing, visualization, and formal analysis. Additionally, TW was responsible for writing the original draft. PD and PD focused on visualization and formal analysis. All authors have read and approved the final manuscript.

CONFLICT OF INTEREST

No conflict of interest.

REFERENCES

Alhakami H, Mirebrahim H, Lonardi S. 2017. A comparative evaluation of genome assembly reconciliation tools. Genome Biology 18: 93.

Arensburger P, Megy K, Waterhouse RM et al. 2010. Sequencing of *Culex quinquefasciatus* establishes a platform for mosquito comparative genomics. Science 330: 86-88.

Bolger A M, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114-2120.

Catapano PL, Falcinelli M, Damiani C, Cappelli A, Koukouli D, Rossi P, Ricci I, Napolioni V, Favia G. 2023. De novo genome assembly of the invasive mosquito species *Aedes japonicus* and *Aedes koreicus*. Parasites and Vectors 16: 427.238

Chaiphongpachara T, Changbunjong T, Laojun S, Nutepsu T, Suwandittakul N, Kuntawong K, Sumruayphol S, Ruangsittichai J. 2022a. Mitochondrial DNA barcoding of mosquito species (Diptera: Culicidae) in Thailand. PLoS One 17: e0275090.

Chaiphongpachara T, Changbunjong T, Sumruayphol S, Laojun S, Suwandittakul N, Kuntawong K. 2022b. Geometric morphometrics versus DNA barcoding for the identification of malaria vectors *Anopheles dirus* and *An. baimaii* in the Thai − Cambodia border. Scientific Reports 12: 13236.

Chaiphongpachara T, Laojun S, Changbunjong T, Sumruayphol S, Suwandittakul N, Chookaew S, Atta Y. 2022c. Genetic diversity, haplotype relationships, and kdr mutation of malaria *Anopheles* vectors in the most *Plasmodium knowlesi*-endemic area of Thailand. Tropical Medicine and Infectious Disease 7: 412.

Chakraborty M, Baldwin-Brown JG, Long A D, Emerson J J. 2016. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Research 44: e147.

Chakraborty M, Ramaiah A, Adolfi A et al. 2020. Hidden features of the malaria vector mosquito, *Anopheles stephensi*, revealed by a high-quality reference genome. bioRxiv 113019.

Dida F, Yi G. 2021. Empirical evaluation of methods for de novo genome assembly. PeerJ Computer Science 7: e636.253

Ellis EA, Storer CG, Kawahara AY. 2021. De novo genome assemblies of butterflies. Gigascience 10: giab041.

Hii J, Rueda LM. 2013. Malaria vectors in the greater Mekong subregion: Overview of malaria vectors and remaining challenges. Southeast Asian Journal of Tropical Medicine and Public Health 44: 73-165

Holt RA, Mani Subramanian G, Halpern A et al. 2002. The genome sequence of the malaria mosquito *Anopheles gambiae.* Science 298: 129-149.

Jiang X, Peery A, Hall AB et al. 2014. Genome analysis of a major urban malaria vector mosquito, *Anopheles stephensi*. Genome Biology 15: 459.

Kaddumukasa MA, Wright J, Muleba M, Stevenson JC, Norris DE, Coetzee M. 2020. Genetic differentiation and population structure of *Anopheles funestus* from Uganda and the southern African countries of Malawi, Mozambique, Zambia and Zimbabwe. Parasites & Vectors 13: 87.

Lau Y L, Lee W C, Chen J, Zhong Z, Jian J, Amir A, Cheong F W, Sum J S, Fong M Y. 2016. Draft genomes of *Anopheles cracens* and *Anopheles maculatus*: Comparison of simian malaria and human malaria vectors in peninsular Malaysia. PLoS One 11: e0157893.

Lischer HEL, Shimizu KK. 2017. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics 18: 474.

Lucas ER, Nagi SC, Egyir-Yawson A et al. 2023. Genome-wide association studies reveal novel loci associated with pyrethroid and organophosphate resistance in *Anopheles gambiae* and *Anopheles coluzzii.* Nature Communications 14: 4946.

Luo R, Liu B, Xie Y et al. 2012. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience 1: 18.

Manni M, Berkeley MR, Seppey M, Zdobnov EM. 2021. BUSCO: Assessing genomic data quality and beyond. Current Protocols 1:e323.

Mwagira-Maina S, Runo S, Wachira L et al. 2021. Genetic markers associated with insecticide

resistance and resting behaviour in *Anopheles gambiae* mosquitoes in selected sites in Kenya. Malaria journal 20: 461.

Nene V, Wortman J R, Lawson D et al. 2007. Genome sequence of *Aedes aegypti*, a major arbovirus vector. Science 316: 1718-1723.

O'Loughlin S M, Okabayashi T, Honda M, Kitazoe Y, Kishino H, Somboon P, Sochantha T, Nambanya S, Saikia P K, Dev V, Walton C. 2008. Complex population history of two *Anopheles dirus* mosquito species in Southeast Asia suggests the influence of Pleistocene climate change rather than human-mediated effects. Journal of Evolutionary Biology 21: 1555-569.

Obsomer V, Defourny P, Coosemans M. 2007. The *Anopheles dirus* complex: Spatial distribution and environmental drivers. Malaria Journal 6: 26.

Parker D M, Carrara V I, Pukrittayakamee S, McGready R, Nosten F H. 2015. Malaria ecology along the Thailand-Myanmar border. Malaria Journal 14: 388.

Rattanarithikul R, Harrison B A, Harbach R E, Panthusiri P, Coleman R E. 2006. Illustrated keys to the mosquitoes of Thailand IV. *Anopheles*. Southeast Asian Journal of Tropical Medicine and Public Health 37: 1-128.

Saeung M, Pengon J, Pethrak C, Thaiudomsap S, Lhaosudto S, Saeung A, Manguin S, Chareonviriyaphap T, Jupatanakul N. 2023. *Dirus* Complex species identification PCR (DiCSIP) improves identification of *Anopheles dirus* complex from Greater Mekong. bioRxiv 561471.

Sarma DK, Prakash A, O'Loughlin SM et al. 2012. Genetic population structure of the malaria vector *Anopheles baimaii* in north-east India using mitochondrial DNA. Malaria journal 11: 76.

Satam H, Joshi K, Mangrolia U et al. 2023. Next-generation sequencing technology: Current trends and advancements. Biology (Basel) 12: 997.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210-3212.

Sumitha M K, Kalimuthu M, Kumar M S et al. 2023. Genetic differentiation among *Aedes aegypti* populations from different eco-geographical zones of India. PLOS Neglected Tropical Diseases 17: e0011486.304

Suwonkerd W, Ritthison W, Ngo CT, Tainchum K, Bangs M J, Chareonviriyaphap T. 2013. Vector biology and malaria transmission in Southeast Asia. *Anopheles* mosquitoes - new insights into malaria vectors.

Tananchai C, Tisgratog R, Juntarajumnong W, Grieco JP, Manguin S, Prabaripai A, Chareonviriyaphap T. 2012. Species diversity and biting activity of *Anopheles dirus* and *Anopheles baimaii* (Diptera: Culicidae) in a malaria prone area of western Thailand. Parasites and Vectors 5: 211.

Vurture G W, Sedlazeck F J, Nattestad M, Underwood C J, Fang H, Gurtowski J, Schatz M C. 2017.

GenomeScope: Fast reference-free genome profiling from short reads. Bioinformatics 33: 2202-2204.

Walton C, Handley J M, Kuvangkadilok C, Collins F H, Harbach R E, Baimai V, Butlin R K. 1999. Identification of five species of the *Anopheles dirus* complex from Thailand, using allele-specific polymerase chain reaction. Medical and Veterinary Entomology 13: 24-32.315

Wang J, Wong G K S, Ni P et al. 2002. RePS: A sequence assembler that masks exact repeats identified from the shotgun data. Genome Research 12: 824-831.

Weedall G D, Riveron J M, Hearn J, Irving H, Kamdem C, Fouet C, White B J, Wondji C S. 2020. An Africa-wide genomic evolution of insecticide resistance in the malaria vector *Anopheles funestus* involves selective sweeps, copy number variations, gene conversion and transposons. PLOS Genetics 16: e1008822.

Zhou D, Zhang D, Ding G et al. 2014. Genome sequence of *Anopheles sinensis* provides insight into genetics basis of mosquito competence for malaria parasites. BMC Genomics 15.

(Manuscript Received: July, 2024; Revised: July, 2024;
Accepted: July, 2024; Online Published: September, 2024)
Online First in www.entosocindia.org and indianentomology.org Ref. No. e24371