Identification and characterization of long noncoding RNAs and their association with acquisition of blood meal in *Culex quinquefasciatus*

Azali Azlan¹, Mardani Abdul Halim¹, Faisal Mohamad¹, Ghows Azzam^{1*}

¹School of Biological Sciences, Universiti Sains Malaysia, 11800 Penang, Malaysia

*Corresponding authors

Email: ghows@usm.my (G.A)

Key words: *Culex quinquefasciatus*, long noncoding RNA, transcriptome, blood-feeding

25 Abstract

26 The Southern house mosquito, *Culex quinquefasciatus* (*Cx. quinquefasciatus*) is an important
 27 vector that transmit multiple diseases including West Nile encephalitis, Japanese encephalitis, St. Louis
 28 encephalitis and lymphatic filariasis. Long noncoding RNAs (lncRNAs) involve in many biological
 29 processes such development, infection, and virus-host interaction. However, there is no systematic
 30 identification and characterization of lncRNAs in *Cx. quinquefasciatus*. Here, we report the first ever
 31 lncRNA identification in *Cx. quinquefasciatus*. By using 31 public RNA-seq datasets, a total of 4,763
 32 novel lncRNA transcripts were identified, of which 3,591, 569, and 603 were intergenic, intronic, and
 33 antisense respectively. Examination of genomic features revealed that *Cx. quinquefasciatus* shared
 34 similar characteristics with other species such as short in length, low GC content, low sequence
 35 conservation, and low coding potential. Furthermore, compared to protein-coding genes, *Cx.*
 36 *quinquefasciatus* lncRNAs had lower expression values, and tended to be expressed in temporally-
 37 specific fashion. In addition, weighted correlation network and functional annotation analyses showed
 38 that lncRNAs may have roles in blood meal acquisition of adult female *Cx. quinquefasciatus*
 39 mosquitoes. This study presents the first systematic identification and analysis of *Cx. quinquefasciatus*
 40 lncRNAs and their association with blood feeding. Results generated from this study will facilitate
 41 future investigation on the function of *Cx. quinquefasciatus* lncRNAs.

42

43

44

45

46

47

48

49

50 Introduction

51 The Southern house mosquito, *Culex quinquefasciatus* (*Cx. quinquefasciatus*), is an important
 52 member of the *Cx. pipiens* complex that spreads multiple types of diseases such as West Nile
 53 encephalitis, Japanese encephalitis, St. Louis encephalitis and lymphatic filariasis (Arensburger et al.,
 54 2010; Nchoutpouen et al., 2019). *Cx. pipiens* complex is composed of six members namely, *Cx.*
 55 *molestus* Forskll, *Cx. pipiens* Linneaus, *Cx. quinquefasciatus* Say, *Cx. pallens* Coquille, *Cx.*
 56 *globocoxitus* Dobrotworsky, and *Cx. australicus* Dobrotworsky & Drummond (Farajollahi et al., 2011;
 57 Zittra et al., 2016). *Cx. quinquefasciatus* mosquitoes are widespread and predominant in urban area
 58 because they are well-adapted to human made habitats (Nchoutpouen et al., 2019). Due to its important
 59 epidemiological role, *Cx. quinquefasciatus* has become a subject for many scientific studies. Several
 60 studies used next-generation sequencing to investigate genome-wide transcriptional response in *Cx.*
 61 *quinquefasciatus* upon blood meal acquisition (Reid et al., 2015; Taparia et al., 2017), insecticide
 62 exposure (Reid et al., 2012; Rezende et al., 2019), virus infections (Göertz et al., 2019; Paradkar et al.,
 63 2015; Rückert et al., 2019, 2019), and vector competence (Shin et al., 2014).

64 Many genes, especially protein-coding, have been implicated to be responsible in blood-
 65 feeding. The ability of *Cx. quinquefasciatus* to carry many harmful diseases is due to its blood-feeding
 66 behavior. Newly eclosed female mosquitoes need a minimum of 48 hours before they can feed on
 67 blood, and subsequently mate (Reid et al., 2015). For example, cytochrome P450s, proteases, and
 68 odorant-binding proteins were upregulated during blood-feeding period (Reid et al., 2015). Another
 69 study performed RNA-seq on antennae of blood-fed and nonblood-fed females, and found that
 70 chemosensory genes, odorant-binding proteins, ionotropic and odorant receptors were among the genes
 71 that were differentially expressed (Taparia et al., 2017). Besides blood meal, RNA-seq was also used to
 72 identify genes that may be involved upon insecticide exposure and resistance. Taken together, the
 73 utilization of RNA-seq approach enables the researcher to observe the differences in *Cx.*
 74 *quinquefasciatus* gene expression level at a genomic scale under different physiological conditions.

Even though protein-coding genes have always been the focus in transcriptomic studies, several reports showed that noncoding RNAs (ncRNAs) are implicated to play major roles in mosquitoes (Azlan et al., 2019b; Etebari et al., 2017, 2016; Göertz et al., 2019; Gu et al., 2013; Miesen et al., 2016; Rückert et al., 2019). Regulatory ncRNAs include small RNAs and long noncoding RNA (lncRNAs). Small RNA consists of micro-RNAs (miRNAs), short-interfering RNAs (siRNAs), and PIWI-interacting RNA (piRNAs), all of which regulate gene expression at the post-transcriptional level (Azlan et al., 2016; Brennecke et al., 2007; Coller and Parker, 2005; Djuranovic et al., 2012). lncRNAs, on the other hand, are RNA species of more than 200 nt in size that do not encode amino acids (Clark and Mattick, 2011; Ulitsky and Bartel, 2013). Due to next-generation sequencing and bioinformatics, identification of lncRNAs has been made possible in many organisms including mosquito such as *Aedes aegypti* (*Ae. aegypti*).

Previous study in *Ae. aegypti* showed that a subset of lncRNAs were highly expressed during early embryo (0-8 hour), and they were also enriched in blood-fed ovary. This suggests that lncRNAs may be maternally inherited, and they may play critical roles in early embryonic development (Azlan et al., 2019b). Beside embryonic development, lncRNAs in mosquitoes are also involved in host-virus interaction. For example, knockdown of *Ae. aegypti* lncRNA (lncRNA_1317) resulted in an increase in virus replication (Etebari et al., 2016). Although these studies suggest the regulatory roles of lncRNAs, the specific mechanism of lncRNAs functions are still unknown. Up until now, there has been no lncRNA studies in *Cx. quinquefasciatus*. We believed that, similar to *Ae. aegypti*, lncRNAs may possess certain regulatory roles in *Cx. quinquefasciatus*. In order to dissect their functional roles, it is important to first systematically annotate lncRNAs in the genome of *Cx. quinquefasciatus*.

Here, we report the first annotation of lncRNAs in *Cx. quinquefasciatus* genome using 31 publicly available RNA-seq data. We applied stringent bioinformatics analysis to confidently predict a total of 4,763 novel lncRNAs that correspond to 4,037 loci in the genome. We then characterized the newly identified lncRNAs such as expression dynamics, genomic features, and sequence conservation.

We also investigate the transcriptional response of *Cx. quinquefasciatus* lncRNAs upon blood meal acquisition. Furthermore, through our co-expression network analysis, a subset of *Cx. quinquefasciatus* lncRNAs may participate in the taking and processing of blood meal.

Materials and Methods

Preparation of public RNA-seq datasets

31 RNA-seq datasets were downloaded from NCBI SRA websites. List of accession number can be found in **Supplementary Data 1**. FASTQC was used to check the quality of raw reads. Trimmomatic version 0.38 (Bolger et al., 2014) was used to clip adapters and low quality reads (<20 phred score).

Alignment of RNA-seq reads to reference genome and transcriptome assembly

Cleaned reads were individually aligned to *Cx. quinquefasciatus* genome (CpipJ2, VectorBase) using Hisat2 (Kim et al., 2015). The reads were mapped according to the strandedness of the libraries. BAM files generated by Hisat2 were used as input for transcriptome assembly using Stringtie version 1.0.1 (Pertea et al., 2015). Annotation file of *Cx. quinquefasciatus* in GTF format (VectorBase) was used to guide the assembly. We set the minimum size of transcripts to be assembled was 200 bp. All Stringtie output files were merged into a single unified transcriptome assembly using Stringtie merge option. The resulting unified transcriptome assembly (GTF format) was compared to the reference annotation using Gffcompare (<https://github.com/gpertea/gffcompare>), and transcripts with class code “u”, “i”, and “x” were kept for lncRNA prediction analysis.

lncRNA prediction

Transcripts with class code “u”, “i”, and “x” were subjected to TransDecoder analysis to predict long ORF, which was set to be minimum of 100 amino acids. Transcripts that do not have long ORF were then analyzed for their coding potential using CPC2 (Kang et al., 2017), PLEK (Li et al., 2014), and

CNCI (Sun et al., 2013). Only transcripts that were categorized as noncoding by all three softwares were retained for downstream analyses. We then used BLASTX to find sequence similarity of the potential lncRNA transcripts against Swissprot database. Transcripts having high sequence similarity (E-value < 1e-3) were discarded. We then removed transcripts that had no strand information and located within less than 2 kb scaffold-end range.

Transcript quantification and expression

Salmon version 0.10.1 (Patro et al., 2017) was used with default parameters to compute the expression of each transcript. We combined protein-coding and lncRNA transcripts together into a single FASTA file when running Salmon. TPM value from Salmon was used for downstream analysis.

Validation of lncRNAs by reverse-transcribed PCR

About 20 adult females of VCRU-lab strain *Cx. quinquefasciatus* were obtained from Vector Control Research Unit, Universiti Sains Malaysia. The mosquitoes were homogenized for RNA isolation and subsequent reverse transcription for cDNA synthesis. For RNA isolation, Rneasy Mini Kit (Qiagen, cat. no: 74104) was used according to manufacturer's protocol. Reverse transcription was carried out using iScript Reverse Transcription Supermix (Bio-Rad Laboratories, California; cat. no. 1708840) according to manufacturer's protocol. We then performed PCR using the cDNA as templates. We designed PCR primers that span exon-exon junction using Primer Blast (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>). Primers used in this study is listed in **Supplementary Data 2**.

Expression specificity

We used JS divergence method to compute expression specificity of lncRNAs and protein-coding genes. MATLAB version R2018b was used to calculate JS score using the formula previously described (Cabili et al., 2011).

150

151 **Co-expression analysis**

152 WGCNA version 1.4.6 was used to perform co-expression analysis between lncRNAs and protein-
153 coding genes (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). We computed variance of
154 each lncRNA and protein-coding gene, and ranked them in descending order. We only picked top 5000
155 genes for co-expression analysis.

156

157 **Functional annotation**

158 g;Profiler was used for functional annotation and enrichment analysis (Reimand et al., 2007). g;SCS
159 threshold was used for multiple testing correction.

160

161 **Result**

162 **Identification of *Cx. quinquefasciatus* lncRNAs**

163 In this study, we used 31 public RNA-seq datasets which were composed of approximately 849
164 million reads as raw inputs for lncRNA prediction. Public RNA-seq datasets used in the study were
165 listed in **Supplementary Data 1**. Pipeline for lncRNA prediction can be found in **Figure 1**. Each
166 library (paired or single-end) was individually mapped against *Cx. quinquefasciatus* genome (CpipJ2,
167 VectorBase) using splice-aware aligner tool, Hisat2 (Kim et al., 2015). Each alignment file was then
168 used for transcriptome assembly using Stringtie, and all 31 Stringtie output files were merged using
169 Stringtie merge option (Pertea et al., 2015). We then compared our merged transcriptome assembly
170 with reference annotation (CpipJ2.4, VectorBase) using gffcompare
171 (<https://github.com/gpertea/gffcompare>). A total of 64,689 transcripts that correspond to 37,532 loci
172 were assembled by Stringtie. Out of 64,689, 45,238 of them were multi-exon transcripts. We managed
173 to recover all *Cx. quinquefasciatus* known transcripts (19,859) using Stringtie; hence, validating our
174 transcriptome assembly.

175 After transcriptome assembly, we then performed stringent transcript filtering steps (**Figure 1**).
 176 Transcripts of less than 200 bp were discarded. We only chose transcripts with class code “i”, “u”, and
 177 “x”, all of which were given by gffcompare. Class code “i”, “u”, and “x” denote intronic, intergenic,
 178 and antisense to reference annotation respectively (<https://github.com/gpertea/gffcompare>). We
 179 discovered that a total of 11,467 transcripts were denoted as either “i”, “u” or “x”, and all of them were
 180 at least 200 bp in length. We then used TransDecoder (Haas et al., 2013) to predict long open-reading
 181 frame (ORF) within these transcripts. Out of 11,467, 2,994 transcripts were discarded because they
 182 were predicted to have ORF of more than 100 amino acids.

183 The remaining 8,473 transcripts were then evaluated for coding potential using CPC2 (Kang et
 184 al., 2017), PLEK (Li et al., 2014), and CNCI (Sun et al., 2013), and those classified as “noncoding” by
 185 all three algorithms were retained for downstream analysis. A total of 7,413 transcripts were found to
 186 be predicted as noncoding by all three softwares. To avoid false positive prediction, we used BLASTX
 187 (E-value cut-off 10^{-3}) to find sequence similarity of those 7,413 transcripts with known proteins in
 188 Swissprot. The number of transcripts that have no significant BLASTX are 7,025. Finally, we removed
 189 unstranded transcripts, and transcripts located within less than 2kb scaffold-end range on the same
 190 strand. By applying this pipeline, we identified a total of 4,763 novel lncRNA transcripts, which
 191 derived from 4,037 distinct loci in the genome. From these 4,763 lncRNA transcripts, 3,591 of them
 192 were intergenic, while the remaining 569 and 603 transcripts were intronic and antisense to reference
 193 gene respectively. List of lncRNAs *Cx. quinquefasciatus* and their corresponding genomic loci can be
 194 found in **Supplementary Data 3**. We randomly selected 5 novel lncRNAs, and validated them through
 195 reverse-transcribed PCR (RT-PCR) using specific primers that span exon-exon junction
 196 (**Supplementary Figure 1**).

197

198 **Structural features of *Cx. quinquefasciatus* lncRNAs**

199 Studies in other species showed that, when compared with protein-coding genes, lncRNAs
200 displayed certain characteristics: shorter in size, lower GC content, low sequence conservation, and
201 lower coding potential (Akay et al., 2019; Azlan et al., 2019b, 2019a; Etebari et al., 2016; Jenkins et al.,
202 2015; Liu et al., 2017; Nam and Bartel, 2012; Wu et al., 2016; Young et al., 2012). We asked if this is
203 also true for *Cx. quinquefasciatus* lncRNAs. We discovered that, in general, *Cx. quinquefasciatus*
204 lncRNAs were significantly shorter than protein-coding transcript (Kolmogorov-Smirnov test (KS-test)
205 p -value < 2.2e-16, **Figure 2a**). Mean of lncRNAs and protein-coding transcripts were 699 bp and 1,377
206 bp respectively. Meanwhile, lncRNAs had a median size of 463 whereas the median length protein-
207 coding transcripts were 1,047 bp.

208 We then checked the GC content of lncRNAs and compared them with protein-coding
209 transcripts, mRNAs. We also included other noncoding sequences such as 3'UTR and
210 5'UTR. We found that, in general, noncoding sequences (lncRNAs, 3'UTR, and 5'UTR) have slightly
211 lower GC content than mRNA (**Figure 2b**). Median GC content of lncRNA, 3'UTR, 5'UTR, and
212 mRNA were 42%, 35%, 45%, and 56% respectively. In addition, mean GC content of noncoding
213 sequences were lower than mRNA (mRNA: 55%, lncRNA: 41%, 5'UTR: 44%, 3'UTR: 36%). Beside
214 GC content, we also examined coding potential of noncoding sequences and compared them with
215 mRNA transcripts. CPC2 software showed that mRNA had the highest median and mean coding
216 probability score (mean: 0.65 and median: 0.672). Median score for lncRNA, 5'UTR, and 3'UTR were
217 0.02, 0.004, and 0.007 respectively. Meanwhile, average probability of lncRNA, 5'UTR and 3'UTR
218 were 0.047, 0.023, and 0.027 respectively.

219 Previous studies reported that lncRNAs have low sequence conservation (Azlan et al., 2019b;
220 Etebari et al., 2016; Wu et al., 2016). Here, we assessed sequence conservation of the newly predicted
221 *Cx. quinquefasciatus* lncRNAs using BLASTN against other insect genomes such as *Ae. aegypti*, *Aedes*
222 *albopictus* (*Ae. albopictus*), *Anopheles gambiae* (*An. gambiae*), and *Drosophila melanogaster* (*D.*
223 *melanogaster*). To determine the level of nucleotide similarity, we used bit score from BLASTN

algorithm (Azlan et al., 2019b; Etebari et al., 2016). Similar to previous reports (Azlan et al., 2019b; Etebari et al., 2016; Wu et al., 2016), *Cx. quinquefasciatus* lncRNAs, in general, had lower sequence conservation than protein-coding transcripts (**Figure 3a**).

***Cx. quinquefasciatus* lncRNAs are lowly expressed and temporally-specific**

To assess the dynamics of lncRNA expression across different time points, we analyzed transcriptome data by Reid et al. (2015), which generated RNA-seq libraries of female mosquitoes *Cx. quinquefasciatus* at 2, 12, 24, 36, 48, 60 and 72 hour post-eclosion (Reid et al., 2015). We reanalyzed this time-course transcriptome data to compute the expression level of *Cx. quinquefasciatus* lncRNAs. We discovered that, across all time points, the expression level of lncRNA was lower than protein-coding genes (KS test p -value $< 2.2e-16$, **Figure 4a**).

Previous studies reported that lncRNA in various organisms displayed a more temporally and tissue-specific pattern in expression (Azlan et al., 2019a, 2019b; Cabili et al., 2011; Wu et al., 2016). We then checked if *Cx. quinquefasciatus* lncRNAs have the same characteristics. To investigate expression specificity of lncRNAs, we employed Jensen-Shannon (JS) score, that ranges between zero to one (Cabili et al., 2011). This specificity score quantifies the similarity in expression value across different time points. Genes having JS score of one indicate the extreme case in which it is only expressed in one time point, whereas those having a score close to zero are ubiquitously expressed in all time points with relatively similar value of expression (Cabili et al., 2011). Similar to previous reports, the expression of *Cx. quinquefasciatus* lncRNAs was more temporally specific than protein-coding genes (KS test p -value $< 2.2e-16$, **Figure 4b**). We found that 58% of lncRNA transcripts had perfect specificity score of one, whereas only 15% of protein-coding transcripts had JS score of one.

***Cx. quinquefasciatus* lncRNAs have potential roles in blood meal acquisition**

248 To investigate the possible involvement of lncRNA in blood meal acquisition, we analyzed
 249 transcriptome data by Reid et al. (2015), which aimed to identify genes, especially protein-coding
 250 genes, that might be crucial for taking and processing the blood meal in adult female *Cx.*
 251 *quinquefasciatus* (Reid et al., 2015). They generated paired-end RNA-seq libraries of seven different
 252 time points after eclosion (2, 12, 24, 36, 48, 60, and 72 hours) of adult females in order to examine
 253 temporal gene expression of post eclosion and prior to blood-feeding (Reid et al., 2015). They showed
 254 that adult females *Cx. quinquefasciatus* need at least 48 hour post-eclosion before they can take their
 255 first blood meal (Reid et al., 2015). Therefore, prior to 48 hours, it was hypothesized that genes
 256 necessary for blood meal acquisition would be differentially expressed, and after 48 hours, the
 257 differentially expressed genes would be important for processing the blood meal. We hypothesized that,
 258 besides protein-coding genes, lncRNAs in *Cx. quinquefasciatus* would also play regulatory roles in
 259 blood-feeding process.

260 Weighted Gene Correlation Network Analysis (WGCNA) has always been used to predict
 261 potential roles of lncRNAs in many species (Azlan et al., 2019a; Langfelder and Horvath, 2008; Wu et
 262 al., 2016; Zhang and Horvath, 2005). WGCNA can be used to search for clusters or modules of highly
 263 correlated genes, which in this case, correlation between lncRNAs and protein-coding genes
 264 (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). To predict the potential roles of lncRNAs
 265 in blood meal acquisition, we used RNA-seq data from Reid et al. 2015, and performed WGCNA
 266 analysis on lncRNAs and protein-coding genes (Reid et al., 2015). WGCNA analysis yielded a total of
 267 21 modules, and 7 of them exhibited strong and statistically significant correlation (correlation > 0.8,
 268 p -value < 0.05) with specific time points (**Figure 5**). Module size ranged from 22 to 611, and the
 269 number of protein-coding genes and lncRNA in each module can be found in Table 1.

270 Two modules were found to be significantly associated with 2 hours post-eclosion, and both of
 271 them had the among the largest number of genes within the clusters (**Table 1**). Functional annotation
 272 using g:Profiler of these two modules revealed that lncRNAs at 2 hours had potential roles in diverse

processes such as compound metabolic pathways, biosynthesis, catabolic process, and proteolysis (Supplemental Data 4). At 12 hour post-eclosion, several genes such as salivary proteins and cytochromes were found to be differentially expressed (Reid et al., 2015). We next asked if the same set of genes with similar functions could be found in our magenta module that was highly associated with 12 hours post-eclosion (correlation: 0.97, p -value: $2e-4$). We found that several protein-coding genes within magenta module were salivary proteins and cytochromes, such as proline-rich salivary peptide (CPIJ007453), threonine-rich salivary mucin (CPIJ010046), cytochrome c oxidase assembly factor (CPIJ010284) and cytochrome P450 (CPIJ011837). Since 43 lncRNAs were clustered together with 59 protein-coding genes within magenta module, we hypothesized that this set of lncRNAs may participate in the same function as protein-coding genes. Therefore, results from WGCNA corroborated with previous findings (Reid et al., 2015), and suggest that lncRNA may intimately function together with protein-coding genes in order to physiologically prepare adult females *Cx. quinquefasciatus* for blood feeding and blood meal processing.

Discussion

In this study, we performed genome-wide identification and characterization of lncRNAs in *Cx. quinquefasciatus* genome using a total of 31 relatively high-depth publicly available RNA-seq libraries. We presented a set of 4,763 novel lncRNAs transcripts which derived from 4,037 loci in *Cx. quinquefasciatus* genome. RNA-seq libraries used here were generated from many tissues and developmental stages, making our prediction pipeline to be relatively robust and comprehensive. We applied stringent filtering steps in our pipeline to reduce false positive and false negative prediction of non-coding transcripts (Azlan et al., 2019a, 2019b; Etebari et al., 2016; Nam and Bartel, 2012; Wu et al., 2016). The filtering steps include using more than one coding potential assessment algorithms and removing transcripts with no strand information (Figure 1). Taken together, lncRNA prediction pipeline used in this study is comparably equivalent to previous studies in other organisms, and it is

298 Analysis of genomic features revealed that lncRNAs identified in our study showed known
 299 characteristics of lncRNAs from other species. The characteristics are shorter in size, low sequence
 300 conservation, and lower GC content. The majority of *Cx. quinquefasciatus* lncRNAs shared high
 301 degree of sequence similarity with *Ae. albopictus* and *Ae. aegypti*, followed by *An. gambiae* and *D.*
 302 *melanogaster*. This lncRNA sequence similarity pattern concur with previous phylogenetic relationship
 303 study on known mosquitoes which showed that *Ae. albopictus* and *Ae. aegypti* are more evolutionarily
 304 closer to *Cx. quinquefasciatus* than *An. gambiae* (Chu et al., 2018). A total of 10 *Cx. quinquefasciatus*
 305 lncRNAs shared high sequence similarity with all insects (**Figure 3b**), suggesting that they are
 306 evolutionarily conserved and may play vital roles in insect development or basic cellular functions.

307 Investigation on the expression across different time points post-eclosion revealed that,
 308 compared to protein-coding genes, *Cx. quinquefasciatus* lncRNAs were more temporally specific and
 309 have lower expression level. This high temporal specificity indicates that lncRNAs have a smaller time
 310 window of expression than protein-coding genes. Even though lncRNAs have lower expression than
 311 protein-coding genes, their temporal specificity in expression suggests that they putatively execute
 312 specific biological functions at specific time point in development.

313 Furthermore, co-expression analysis by WCGNA revealed that *Cx. quinquefasciatus* lncRNAs
 314 were significantly correlated with specific time-point of pre-blood feeding immediately after eclosion,
 315 suggesting that lncRNAs may possess potential roles in blood meal acquisition by adult female
 316 mosquitoes. This blood meal acquisition by adult female is physiologically important as it is not only
 317 required for reproduction and mating, but it serves as a gateway for pathogens, making *Cx.*
 318 *quinquefasciatus* one of the most harmful mosquito vectors on Earth. However, the data presented here
 319 is simply descriptive, and it does not provide empirical evidence via experimentation on the specific
 320 mechanisms of lncRNA functions in blood meal acquisition. We believed that results generated in this
 321 study can be a starting point for dissecting the mechanism of *Cx. quinquefasciatus* lncRNAs functions.
 322 In conclusion, this research provides the first comprehensive annotation and characterization of *Cx.*

323 *quinquefasciatus* lncRNAs, and their putative roles in blood-feeding. We believed that data presented
324 here will be a valuable resource for genetics and molecular studies of ncRNAs in *Cx. quinquefasciatus*.

325

326 **Acknowledgment**

327 We would like to thank all our collaborators and colleagues for the discussion and the work conducted
328 in this lab. We would especially like to thank the Vector Control Research Unit, Universiti Sains
329 Malaysia for providing the mosquito samples. This study was funded by the Universiti Sains Malaysia
330 Research University Grant (1001/PBIOLOGI/811320) and MESTECC ScienceFund
331 (305/PBIOLOGI/613238).

332

333 **Figure Legend**

334 **Figure 1: lncRNA identification**

335 Summary of lncRNA prediction pipeline

336

337 **Figure 2: Features of *Cx. quinquefasciatus* lncRNAs**

338 a) Distribution of transcript length b) GC content c) Coding potential from CPC2 algorithm

339

340 **Figure 3: Sequence conservation of *Cx. quinquefasciatus* lncRNAs**

341 a) Boxplot of similarity bit score using BLASTN algorithm. b) The Venn diagram shows the number of
342 *Cx. quinquefasciatus* with E-value below 1e-5 in other species using BLASTN algorithm.

343

344 **Figure 4: Expression of *Cx. quinquefasciatus* lncRNAs**

345 a) Distribution of expression of lncRNAs and protein-coding genes across different time points. b)
346 Distribution of specificity score of lncRNAs and protein-coding genes.

347

348 **Figure 5: Module-trait correlation**

349 Module-trait correlation and corresponding p-values was computed using WGCNA package. Red
350 indicated positive correlation (correlation > 0) whereas blue color indicated negative correlation
351 (correlation < 0). PE: post-eclosion.

352

353 **Table 1: Number of lncRNA and protein-coding genes in each module (Correlation > 0.8, P-value**
354 **< 0.05)**

355 **Supplementary Figure 1: PCR products in 1.5% agarose gel**

356 **Supplementary Data 1: List of 31 publicly available RNA-seq libraries used in this study**

357 **Supplementary Data 2: Primers used in this study**

358 **Supplementary Data 3: List of lncRNAs in *Cx. quinquefasciatus* genome**

359 **Supplementary Data 4: Gene ontology and enrichment from g:Profiler**

360

361

362 **References**

- Akay, A., Jordan, D., Navarro, I.C., Wrzesinski, T., Ponting, C.P., Miska, E.A., Haerty, W., 2019. Identification of functional long non-coding RNAs in *C. elegans*. *BMC Biol.* 17. <https://doi.org/10.1186/s12915-019-0635-7>
- Arensburger, P., Megy, K., Waterhouse, R.M., Abrudan, J., Amedeo, P., Antelo, B., Bartholomay, L., Bidwell, S., Caler, E., Camara, F., Campbell, C.L., Campbell, K.S., Casola, C., Castro, M.T., Chandramouliswaran, I., Chapman, S.B., Christley, S., Costas, J., Eisenstadt, E., Feschotte, C., Fraser-Liggett, C., Guigo, R., Haas, B., Hammond, M., Hansson, B.S., Hemingway, J., Hill, S.R., Howarth, C., Ignell, R., Kennedy, R.C., Kodira, C.D., Lobo, N.F., Mao, C., Mayhew, G., Michel, K., Mori, A., Liu, N., Naveira, H., Nene, V., Nguyen, N., Pearson, M.D., Pritham, E.J., Puiu, D., Qi, Y., Ranson, H., Ribeiro, J.M.C., Roberston, H.M., Severson, D.W., Shumway, M., Stanke, M., Strausberg, R.L., Sun, C., Sutton, G., Tu, Z. (Jake), Tubio, J.M.C., Unger, M.F., Vanlandingham, D.L., Vilella, A.J., White, O., White, J.R., Wondji, C.S., Wortman, J., Zdobnov, E.M., Birren, B., Christensen, B.M., Collins, F.H., Cornel, A., Dimopoulos, G., Hannick, L.I., Higgs, S., Lanzaro, G.C., Lawson, D., Lee, N.H., Muskavitch, M.A.T., Raikhel, A.S., Atkinson, P.W., 2010. Sequencing of *Culex quinquefasciatus* Establishes a Platform for Mosquito Comparative Genomics. *Science* 330, 86–88. <https://doi.org/10.1126/science.1191864>
- Azlan, A., Dzaki, N., Azzam, G., 2016. Argonaute: The executor of small RNA function. *J. Genet. Genomics* 43, 481–494. <https://doi.org/10.1016/j.jgg.2016.06.002>

- Azlan, A., Halim, M.A., Azzam, G., 2019a. Genome-wide identification and characterization of long intergenic noncoding RNAs in the regenerative flatworm *Macrostomum lignano*. *Genomics*. <https://doi.org/10.1016/j.ygeno.2019.07.016>
- Azlan, A., Obeidat, S.M., Yunus, M.A., Azzam, G., 2019b. Systematic identification and characterization of *Aedes aegypti* long noncoding RNAs (lncRNAs). *Sci. Rep.* 9. <https://doi.org/10.1038/s41598-019-47506-9>
- Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma. Oxf. Engl.* 30, 2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Brennecke, J., Aravin, A.A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., Hannon, G.J., 2007. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in *Drosophila*. *Cell* 128, 1089–1103. <https://doi.org/10.1016/j.cell.2007.01.043>
- Cabili, M.N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A., Rinn, J.L., 2011. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev.* 25, 1915–1927. <https://doi.org/10.1101/gad.1744661>
- Chu, H., Li, C., Guo, X., Zhang, H., Luo, P., Wu, Z., Wang, G., Zhao, T., 2018. The phylogenetic relationships of known mosquito (Diptera: Culicidae) mitogenomes. *Mitochondrial DNA Part A* 29, 31–35. <https://doi.org/10.1080/24701394.2016.1233533>
- Clark, M.B., Mattick, J.S., 2011. Long noncoding RNAs in cell biology. *Semin. Cell Dev. Biol.* 22, 366–376. <https://doi.org/10.1016/j.semedb.2011.01.001>
- Coller, J., Parker, R., 2005. General Translational Repression by Activators of mRNA Decapping. *Cell* 122, 875–886. <https://doi.org/10.1016/j.cell.2005.07.012>
- Djuranovic, S., Nahvi, A., Green, R., 2012. miRNA-Mediated Gene Silencing by Translational Repression Followed by mRNA Deadenylation and Decay. *Science* 336, 237–240. <https://doi.org/10.1126/science.1215691>
- Etebari, K., Asad, S., Zhang, G., Asgari, S., 2016. Identification of *Aedes aegypti* Long Intergenic Non-coding RNAs and Their Association with Wolbachia and Dengue Virus Infection. *PLoS Negl. Trop. Dis.* 10, e0005069. <https://doi.org/10.1371/journal.pntd.0005069>
- Etebari, K., Hegde, S., Saldaña, M.A., Widen, S.G., Wood, T.G., Asgari, S., Hughes, G.L., 2017. Global Transcriptome Analysis of *Aedes aegypti* Mosquitoes in Response to Zika Virus Infection. *mSphere* 2, e00456-17. <https://doi.org/10.1128/mSphere.00456-17>
- Farajollahi, A., Fonseca, D.M., Kramer, L.D., Marm Kilpatrick, A., 2011. “Bird biting” mosquitoes and human disease: A review of the role of *Culex pipiens* complex mosquitoes in epidemiology. *Infect. Genet. Evol.* 11, 1577–1585. <https://doi.org/10.1016/j.meegid.2011.08.013>
- Göertz, G.P., Miesen, P., Overheul, G.J., van Rij, R.P., van Oers, M.M., Pijlman, G.P., 2019. Mosquito Small RNA Responses to West Nile and Insect-Specific Virus Infections in *Aedes* and *Culex* Mosquito Cells. *Viruses* 11. <https://doi.org/10.3390/v11030271>
- Gu, J., Hu, W., Wu, J., Zheng, P., Chen, M., James, A.A., Chen, X., Tu, Z., 2013. miRNA Genes of an Invasive Vector Mosquito, *Aedes albopictus*. *PLOS ONE* 8, e67638. <https://doi.org/10.1371/journal.pone.0067638>
- Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. *Nat. Protoc.* 8. <https://doi.org/10.1038/nprot.2013.084>
- Jenkins, A.M., Waterhouse, R.M., Muskavitch, M.A., 2015. Long non-coding RNA discovery across the genus *Anopheles* reveals conserved secondary structures within and beyond the *Gambiae* complex. *BMC Genomics* 16, 337. <https://doi.org/10.1186/s12864-015-1507-3>

- Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., Gao, G., 2017. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. *Nucleic Acids Res.* 45, W12–W16. <https://doi.org/10.1093/nar/gkx428>
- Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* 12, 357–360. <https://doi.org/10.1038/nmeth.3317>
- Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9, 559. <https://doi.org/10.1186/1471-2105-9-559>
- Li, A., Zhang, J., Zhou, Z., 2014. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics* 15, 311. <https://doi.org/10.1186/1471-2105-15-311>
- Liu, F., Guo, D., Yuan, Z., Chen, C., Xiao, H., 2017. Genome-wide identification of long non-coding RNA genes and their association with insecticide resistance and metamorphosis in diamondback moth, *Plutella xylostella*. *Sci. Rep.* 7. <https://doi.org/10.1038/s41598-017-16057-2>
- Miesen, P., Ivens, A., Buck, A.H., Rij, R.P. van, 2016. Small RNA Profiling in Dengue Virus 2-Infected Aedes Mosquito Cells Reveals Viral piRNAs and Novel Host miRNAs. *PLoS Negl. Trop. Dis.* 10, e0004452. <https://doi.org/10.1371/journal.pntd.0004452>
- Nam, J.-W., Bartel, D.P., 2012. Long noncoding RNAs in *C. elegans*. *Genome Res.* 22, 2529–2540. <https://doi.org/10.1101/gr.140475.112>
- Nchoutpouen, E., Talipouo, A., Djiappi-Tchamen, B., Djamouko-Djonkam, L., Kopya, E., Ngadjeu, C.S., Doumbe-Belisse, P., Awono-Ambene, P., Kekeunou, S., Wondji, C.S., Antonio-Nkondjio, C., 2019. Culex species diversity, susceptibility to insecticides and role as potential vector of Lymphatic filariasis in the city of Yaoundé, Cameroon. *PLoS Negl. Trop. Dis.* 13, e0007229. <https://doi.org/10.1371/journal.pntd.0007229>
- Paradkar, P.N., Duchemin, J.-B., Rodriguez-Andres, J., Trinidad, L., Walker, P.J., 2015. Cullin4 Is Pro-Viral during West Nile Virus Infection of Culex Mosquitoes. *PLOS Pathog.* 11, e1005143. <https://doi.org/10.1371/journal.ppat.1005143>
- Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat. Methods* 14, 417–419. <https://doi.org/10.1038/nmeth.4197>
- Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. <https://doi.org/10.1038/nbt.3122>
- Reid, W.R., Zhang, L., Liu, F., Liu, N., 2012. The Transcriptome Profile of the Mosquito Culex quinquefasciatus following Permethrin Selection. *PLOS ONE* 7, e47163. <https://doi.org/10.1371/journal.pone.0047163>
- Reid, W.R., Zhang, L., Liu, N., 2015. Temporal Gene Expression Profiles of Pre Blood-Fed Adult Females Immediately Following Eclosion in the Southern House Mosquito Culex Quinquefasciatus. *Int. J. Biol. Sci.* 11, 1306–1313. <https://doi.org/10.7150/ijbs.12829>
- Reimand, J., Kull, M., Peterson, H., Hansen, J., Vilo, J., 2007. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.* 35, W193–W200. <https://doi.org/10.1093/nar/gkm226>
- Rezende, T.M.T., Rezende, A.M., Luz Wallau, G., Santos Vasconcelos, C.R., de-Melo-Neto, O.P., Silva-Filha, M.H.N.L., Romão, T.P., 2019. A differential transcriptional profile by Culex quinquefasciatus larvae resistant to Lysinibacillus sphaericus IAB59 highlights genes and pathways associated with the resistance phenotype. *Parasit. Vectors* 12, 407. <https://doi.org/10.1186/s13071-019-3661-y>
- Rückert, C., Prasad, A.N., Garcia-Luna, S.M., Robison, A., Grubaugh, N.D., Weger-Lucarelli, J., Ebel, G.D., 2019. Small RNA responses of Culex mosquitoes and cell lines during acute and

- persistent virus infection. *Insect Biochem. Mol. Biol.* 109, 13–23.
<https://doi.org/10.1016/j.ibmb.2019.04.008>
- Shin, D., Civana, A., Acevedo, C., Smartt, C.T., 2014. Transcriptomics of differential vector competence: West Nile virus infection in two populations of *Culex pipiens quinquefasciatus* linked to ovary development. *BMC Genomics* 15, 513. <https://doi.org/10.1186/1471-2164-15-513>
- Sun, L., Luo, H., Bu, D., Zhao, G., Yu, K., Zhang, C., Liu, Y., Chen, R., Zhao, Y., 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Res.* 41, e166–e166. <https://doi.org/10.1093/nar/gkt646>
- Taparia, T., Ignell, R., Hill, S.R., 2017. Blood meal induced regulation of the chemosensory gene repertoire in the southern house mosquito. *BMC Genomics* 18, 393.
<https://doi.org/10.1186/s12864-017-3779-2>
- Ulitsky, I., Bartel, D.P., 2013. lincRNAs: Genomics, Evolution, and Mechanisms. *Cell* 154, 26–46.
<https://doi.org/10.1016/j.cell.2013.06.020>
- Wu, Y., Cheng, T., Liu, C., Liu, D., Zhang, Q., Long, R., Zhao, P., Xia, Q., 2016. Systematic Identification and Characterization of Long Non-Coding RNAs in the Silkworm, *Bombyx mori*. *PLOS ONE* 11, e0147147. <https://doi.org/10.1371/journal.pone.0147147>
- Young, R.S., Marques, A.C., Tibbit, C., Haerty, W., Bassett, A.R., Liu, J.-L., Ponting, C.P., 2012. Identification and Properties of 1,119 Candidate LincRNA Loci in the *Drosophila melanogaster* Genome. *Genome Biol. Evol.* 4, 427–442. <https://doi.org/10.1093/gbe/evs020>
- Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* 4, Article17. <https://doi.org/10.2202/1544-6115.1128>
- Zittra, C., Flechl, E., Kothmayer, M., Vitecek, S., Rossiter, H., Zechmeister, T., Fuehrer, H.-P., 2016. Ecological characterization and molecular differentiation of *Culex pipiens* complex taxa and *Culex torrentium* in eastern Austria. *Parasit. Vectors* 9, 197. <https://doi.org/10.1186/s13071-016-1495-4>

363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385

386
387
388
389

390

391 **Table 1: Number of lncRNA and protein-coding genes in each module (Correlation > 0.8, P-value**
392 **< 0.05)**

393

394

Modules	lncRNA	Protein-coding	Hour post-eclosion	Correlation (P-value)
Green	66	226	2 hours	0.81 (0.03)
Brown	103	508	2 hours	0.99 (2e-5)
Magenta	43	59	12 hours	0.97 (2e-4)
Black	41	150	24 hours	0.84 (0.02)
Royal blue	17	7	36 hours	0.98 (6e-5)
Light yellow	16	13	48 hours	0.99 (1e-5)
Dark red	12	10	60 hours	0.99 (4e-5)

395

Figure 1: lncRNA identification

Summary of lncRNA prediction pipeline

Figure 2: Features of *Cx. quinquefasciatus* lncRNAs

a) Distribution of transcript length b) GC content c) Coding potential from CPC2 algorithm

Figure 3: Sequence conservation of *Cx. quinquefasciatus* lncRNAs

a) Boxplot of similarity bit score using BLASTN algorithm. b) The Venn diagram shows the number of *Cx. quinquefasciatus* with E-value below $1e-5$ in other species using BLASTN algorithm.

a

b

Figure 4: Expression of *Cx. quinquefasciatus* lncRNAs

a) Distribution of expression of lncRNAs and protein-coding genes across different time points. b) Distribution of specificity score of lncRNAs and protein-coding genes.

Module-trait relationships

Figure 5: Module-trait correlation
Module-trait correlation and corresponding p-values was computed using WGCNA package. Red indicated positive correlation (correlation > 0) whereas blue color indicated negative correlation (correlation < 0). PE: post-eclosion.

Supplementary Figure 1: PCR products in 1.5% agarose gel

Supplementary Data 1: List of 31 publicly available RNA-seq libraries used in this study

Accession number	Condition	Single/Paired	Reference/Publication
SRR5500831	sugar-fed replicate 2	Single	Taparia et al., 2017
SRR5500832	sugar-fed replicate 1	Single	
SRR5500833	blood-fed replicate 2	Single	
SRR5500834	blood-fed replicate 1	Single	
SRR1005577	adult female 2 hour post-eclosion	Paired	Reid et al., 2015
SRR1005578	adult female 12 hour post-eclosion	Paired	
SRR1005579	adult female 24 hour post-eclosion	Paired	
SRR1005580	adult female 36 hour post-eclosion	Paired	
SRR1005581	adult female 48 hour post-eclosion	Paired	
SRR1005582	adult female 60 hour post-eclosion	Paired	
SRR1005583	adult female 72 hour post-eclosion	Paired	
SRR1539388	control uninfected cell line	Paired	Paradkar et al., 2015
SRR1539389	control uninfected cell line	Paired	
SRR1539390	cell infected with West Nile virus	Paired	
SRR1539391	cell infected with West Nile virus	Paired	
SRR9617919	Susceptible colony	Paired	Rezende et al., 2019
SRR9617921	Susceptible colony	Paired	
SRR9617922	Susceptible colony	Paired	
SRR9617920	Resistant colony	Paired	
SRR9617923	Resistant colony	Paired	
SRR9617924	Resistant colony	Paired	
SRR407541	Egg	Single	-
SRR407542	Third instar larvae	Single	
SRR407543	Pupae	Single	
SRR407545	Adult	Single	
SRR407546	S-lab strain third instar larvae	Single	
SRR1050976	Adult female	Paired	-
SRR1050983	Adult female	Paired	
SRR1462324	Adult female	Paired	-
SRR991016	Female antennae	Paired	-
SRR991017	Female hind leg	Paired	

Supplementary Data 2: Primers used in this study

Primers	Sequence	Size of PCR product (bp)
lncRNA_88.1 forward	ACAACACCGTTTGGACAAGT	155
lncRNA_88.1 reverse	ACAATCCAATCATTTCCCCGT	
lncRNA_8741.11 forward	TCGGCTCACGTGAATCTGAAA	173
lncRNA_8741.11 reverse	TGGAGCTGATTAGGCTTGCT	
lncRNA_9150.1 forward	AACGATGGGTCAGATGGTGG	109
lncRNA_9150.1 reverse	CCCTGTCTCGAGATATGCCC	
lncRNA_9394.4 forward	ACAGGTTGCCTCTTGTTCCT	139
lncRNA_9394.4 reverse	GTGCGACTCGTGCTGAAAAA	
lncRNA_9409.1 forward	TCGATCAGTTTGTGGGGCAA	108
lncRNA_9409.1 reverse	CTGCGATCTGGATTTGCAGC	