OMICS A Journal of Integrative Biology
Volume 21, Number 8, 2017
© Mary Ann Liebert, Inc.
DOI: 10.1089/omi.2017.0073

Research Article

Mosquito-Borne Diseases and Omics: Tissue-Restricted Expression and Alternative Splicing Revealed by Transcriptome Profiling of *Anopheles stephensi*

Sreelakshmi K. Sreenivasamurthy,[1,2] Anil K. Madugundu,[1,3] Arun H. Patil,[1,4,5] Gourav Dey,[1,2]
Ajeet Kumar Mohanty,[6,7] Manish Kumar,[1,2] Krishna Patel,[1,8] Charles Wang,[9] Ashwani Kumar,[6]
Akhilesh Pandey,[1,10–13] and Thottethodi Subrahmanya Keshava Prasad[1,4]

Abstract

Malaria is one of the most debilitating mosquito-borne diseases with high global health burdens. While much of the research on malaria and mosquito-borne diseases is focused on Africa, Southeast Asia accounts for a sizable portion of the global burden of malaria. Moreover, about 50% of the Asian malaria incidence and deaths have been from India. A promising development in this context is that the completion of genome sequence of *Anopheles stephensi*, a major malaria vector in Asia, offers new opportunities for global health innovation, including the progress in deciphering the vectorial ability of this mosquito species at a molecular level. Moving forward, tissue-based expression profiling would be the next obvious step in understanding gene functions of *An. stephensi*. We report in this article, to the best of our knowledge, the first in-depth study on tissue-based transcriptomic profile of four important organs (midgut, Malpighian tubules, fat body, and ovary) of adult female *An. stephensi* mosquitoes. In all, we identified over 20,000 transcripts corresponding to more than 12,000 gene loci from these four tissues. We present and discuss the tissue-based expression profiles of majority of annotated transcripts in *An. stephensi* genome, and the dynamics of their alternative splicing in these tissues, in this study. The domain-based Gene Ontology analysis of the differentially expressed transcripts in each of the mosquito tissue indicated enrichment of transcripts with proteolytic activity in midgut; transporter activity in Malpighian tubules; cell cycle, DNA replication, and repair activities in ovaries; and oxidoreductase activities in fat body. Tissue-based study of transcript expression and gene functions markedly enhances our understanding of this important malaria vector, and in turn, offers rationales for further studies on vectorial ability and identification of novel molecular targets to intercept malaria transmission.

Keywords: *Anopheles stephensi*, global health innovation, malaria transmission, mosquito-borne diseases, transcriptomics

Introduction

Malaria poses a significant challenge in the tropical and subtropical regions as a life-threatening mosquito-borne disease. According to World Health Organization (WHO) World Malaria Report in 2016, there were ~212 million malaria cases in the year 2015, which resulted in an estimated death of about 429,000 individuals globally. Most of these cases (90%) are from the African region. Southeast Asia accounts for about 7% incidence. About 50%

[1]Institute of Bioinformatics, Bangalore, India.
[2]Manipal University, Manipal, India.
[3]Centre for Bioinformatics, Pondicherry University, Kalapet, India.
[4]YU-IOB Center for Systems Biology and Molecular Medicine, Yenepoya University, Mangalore, India.
[5]School of Biotechnology, KIIT University, Bhubaneswar, India.
[6]National Institute of Malaria Research, Field Station, Panjim, India.
[7]Department of Zoology, Goa University, Taleigao Plateau, India.
[8]Amrita School of Biotechnology, Amrita Vishwa Vidyapeetham, Kollam, India.
[9]Center for Genomics and Department of Basic Sciences, School of Medicine, Loma Linda University, Loma Linda, California.
[10]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland.
Departments of [11]Biological Chemistry, [12]Oncology, and [13]Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland.

of the Asian malaria incidence and deaths have been from India (World Health Organization, 2016). The latest updates on the cases and deaths reported in India are monitored by the National Vector Borne Disease Control Programme (NVBDCP), according to which there has been a decline in the number of malaria cases, with about a million cases of malaria reported in the year 2014 (www.nvbdcp.gov.in/malaria3.html). Control of ''*Anopheles*'' mosquito vectors has been crucial in the control of malaria transmission. Out of the 41 different *Anopheline* species reported as significant vectors for transmission of human malaria, *Anopheles stephensi* is an important vector in India and South Asia (Kumar et al., 2016; Sinka et al., 2012).

Sequencing of *Anopheles gambiae* genome in the year 2002 not only led to an increase in the molecular-level investigations of *An. gambiae* but also opened new avenues in other species by orthology and sequencing-based studies. The trend is depicted by a PubMed search with the keyword ''Anopheles'' that resulted in 14,576 publications, majority of which have been after the year 2000 as shown in the Supplementary Figure S1. Most of the studies of postgenome sequencing have been focused on the role of various genes and development of methods to regulate their expression.

The overall aim of the community has been to embark on a feasible means to control the spread of infectious organisms either by controlling the vector/mosquito population or by curbing their vectorial ability. Although, numerous studies have already been performed, in this regard, on the recently sequenced malarial vectors (Childs et al., 2016; Gantz et al., 2015; Mitchell et al., 2015; Venkatrao et al., 2017), the focus has been on previously studied molecules than new targets. This could probably be due to the lack of reliable data owing to incomplete genome assemblies and annotation in the identification of such targets.

Evolution and adaptation of the malarial vector species in atypical habitats and their behavioral changes have further impaired efforts on vector control (Ramasamy and Surendran, 2016; Sougoufara et al., 2017). Therefore, newer and better avenues are being sought for control of these vectors. Toward this, we followed an integrated OMICs approach in the revision of genome assembly and annotation of these malarial vectors using proteomic and transcriptomic data (Chaerkady et al., 2011; Prasad et al., 2017). Although transcriptomic data played a key role in refining the assembly and annotation of the genomes of *An. stephensi* in the previous study, a *tissue-based* expression profile was not thoroughly analyzed.

In this study, we report our efforts toward a global comparative transcriptome analysis, including the identification of spliced variants of the four important tissues of female *An. stephensi*—midgut, ovary, Malpighian tubule, and fat body.

Materials and Methods

RNA isolation and sequencing

Adult female *An. stephensi* mosquitoes grown at the National Institute of Malaria Research (NIMR Field Station (Panjim, India) were dissected to obtain midgut, Malpighian tubules, ovaries, and fat body. These dissected tissues were stored in RNA later to preserve the RNA quality till RNA extraction. The RNA isolation and sequencing were performed as described earlier (Kelkar et al., 2014; Prasad et al., 2017). In brief, the RNA isolated using Qiagen miRNeasy Kit was used for the preparation of indexed RNA-seq libraries using TruSeq RNA Sample Preparation and SBS Kit v3. The indexed and pooled libraries were sequenced on two lanes (as technical replicates) of Illumina HiScan SQ platform. The study was reviewed and approved by the Scientific Review Board of National Institute of Malaria Research (ICMR), New Delhi (NIMR/IDVC/2010/54, December 2010) and the Institutional BioSafety Committee of Institute of Bioinformatics (BT/BS/17/93/2006-PID, May 2013).

Read alignment and transcript assembly

The raw reads were filtered based on base quality to remove ambiguous bases present due to the sequencing errors at the 3′ end of the reads. Base quality filter of ≥20 was considered as good and accepted. FastQC (Version 0.10.1) tool was used to determine the quality of the raw data. Poor-quality calls with Phred score <20 were trimmed off using fqtrim v0.9.4. Post-trimming, reads that were less than 60 bp in length were also discarded (about 6%), to avoid ambiguous alignment. Quality-filtered reads were aligned against *An. stephensi* genome build (ASTEI2) downloaded from VectorBase (www.vectorbase.org) using HISAT (Version 2.1.0) (Kim et al., 2015) aligner with the default parameters. HISAT2 was supplied with known annotations and Gene Transfer File (GTF), AsteI2.2 from VectorBase. The alignment of reads from each lane for each tissue was carried out individually against the reference genome resulting in eight different ''Binary Alignment Map'' (BAM) files.

The BAM files for each tissue were then merged to obtain merged BAM files, one for each tissue. The aligned reads were assembled against the AsteI2.2 gene annotations, as reference, using the StringTie (version 1.2.1) assembler (Pertea et al., 2015). Assembled transcripts were further quantified and annotated into known and novel categories using the ''gffcompare'' in StringTie package as described earlier (Pertea et al., 2016). To determine all expressed transcripts as a GTF file, StringTie assemblies built for four tissues were merged using StringTie-merge option. Novel isoforms and intergenic transcripts were obtained by comparing the merged StringTie assemblies of all the four tissues to the annotated transcripts from VectorBase using gffcompare.

Coding potential of the identified transcripts was predicted using the Coding Potential Assessment Tool (CPAT) version 1.2.2 (Wang et al., 2013). Transcripts which were >200 bp in length with a CPAT score threshold of <0.39 were categorized as long noncoding RNAs (lncRNAs) according to the prebuilt classification model for the fly genes.

Identification of differentially expressed genes across four tissues

Merged GTF file from StringTie was annotated into different classes of transcripts using gffcompare with respect to the VectorBase annotations. Expression levels of transcripts as determined by the StringTie assembler were compared across tissues. The expression information from individual lanes was used as technical replicates for each tissue. Differential expression was computed using Cuffdiff after normalizing the data across samples by calculating Fragments per Kilobase of exon per Million Fragments Mapped (FPKM) (Trapnell et al., 2013). The R-package version 2.18.0 of cummeRbund was used for visualization, analysis of RNA-seq data, and cluster generation (Goff et al., 2013). An overview of the analysis pipeline is provided in Figure 1. To identify tissue-specific transcripts, we initially filtered transcripts with FPKM value ≥1.0 in at least one among the four tissue types. We then applied the right-tailed *t*-test to identify the transcripts, which are significantly abundant in one tissue as against other tissues.

Results

Transcriptome sequencing of four *An. stephensi* tissues—midgut, Malpighian tubules, fat body, and ovary, was performed to create a tissue-based unbiased expression profile of vector genes. In total, about 500 million 100 bp long paired-end reads were generated from four tissues, with about 55 million read pairs per tissue sample. The expression levels of transcripts between the replicates and among the tissues were compared. Figure 2A represents the variations between intertissue and intratissue transcript expression in the form of a distance-based heatmap. As expected, variations were found to be minimal between the replicates and relatively higher between the tissues, with ovary and Malpighian tubules being the most diverse.

By following the standard alignment and assembly pipeline using the HISAT2 and StringTie assembler, we identified a total of 25,795 transcripts. However, after the initial filtering for the FPKM values (≥0.1), we retained only 23,009 transcripts corresponding to 12,256 gene loci. Table 1 presents the summary of number of transcripts identified across different classes in the four tissues with their corresponding gene loci. The expression of these transcripts was comparable across tissues with the median FPKM value ranging about 2–3 in all the tissues as represented by the box plot in Figure 2B. Figure 2C and 2D provides the general distribution of the length and the FPKM values of the transcript assemblies across the four tissues. About 60% of the transcripts were found to have FPKM value of 1 and above.

The average length of majority of the transcripts tends to be in the range of 1000–3000 bp. This demonstrates an expected trend of a reliable depth and absence of any skewness. The transcript assemblies were classified into different classes using gffcompare. However, to avoid overinterpretation of the data, we have only focused our findings on the known ''=,'' alternate ''j,'' and intergenic unknown ''u'' class of the transcript assemblies for our analysis. Supplementary Table S1 provides the list of transcripts classified based on the three class codes along with the FPKM values for each of the transcripts identified.

In our analysis, we noticed that almost equivalent number of transcript assemblies were segregated into the known (=) and the alternate (j) categories. A deeper analysis of the annotated transcripts showed that almost all the genes in the VectorBase genome had a single transcript annotated. Further manual inspection into this matter revealed that the untranslated regions (UTRs) of predicted transcript models in the current annotation are missed, probably due to the poor quality of currently available data for this strain.

UTR analysis of the identified transcripts that were classified as alternative transcripts revealed that 2,265 of them differed from their corresponding annotated transcript only in their UTR regions (Supplementary Table S2). The length of the coding sequence (CDS) remained the same as the annotated/known transcript for these 2,265 alternative transcripts. This leads to the possibility that the transcript assemblies with the extension of the exonic regions supported by the reads were classified as alternate transcripts instead of real UTRs of the reference transcripts that were missed during prediction.

Tissue restricted transcripts

Majority of the transcripts identified (about 87%) were found to be expressed largely at similar levels in all the four tissues and the other 13% of the transcripts identified seemed

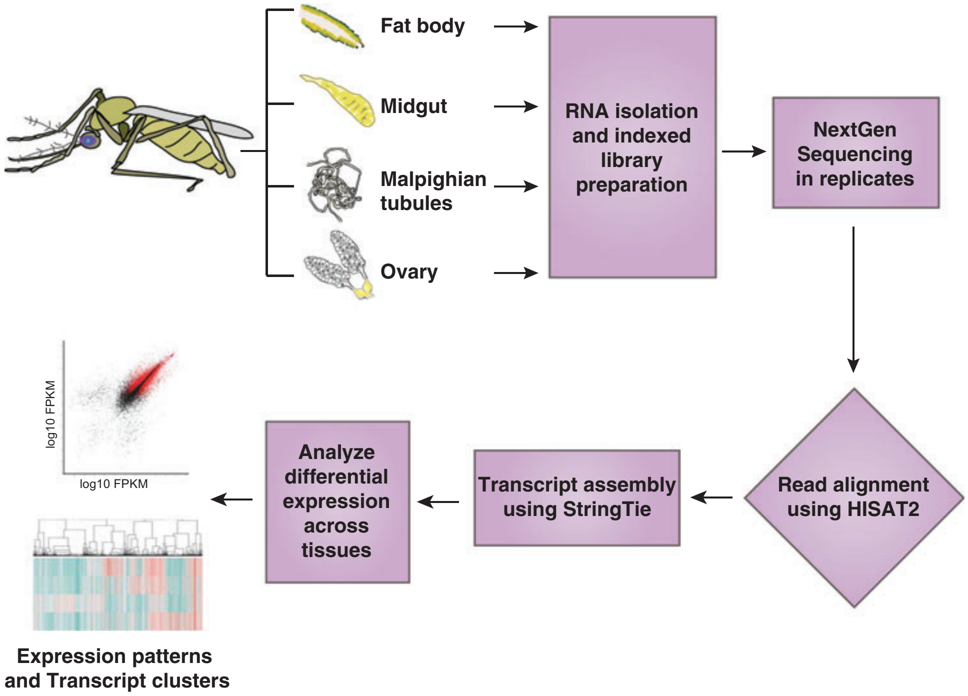
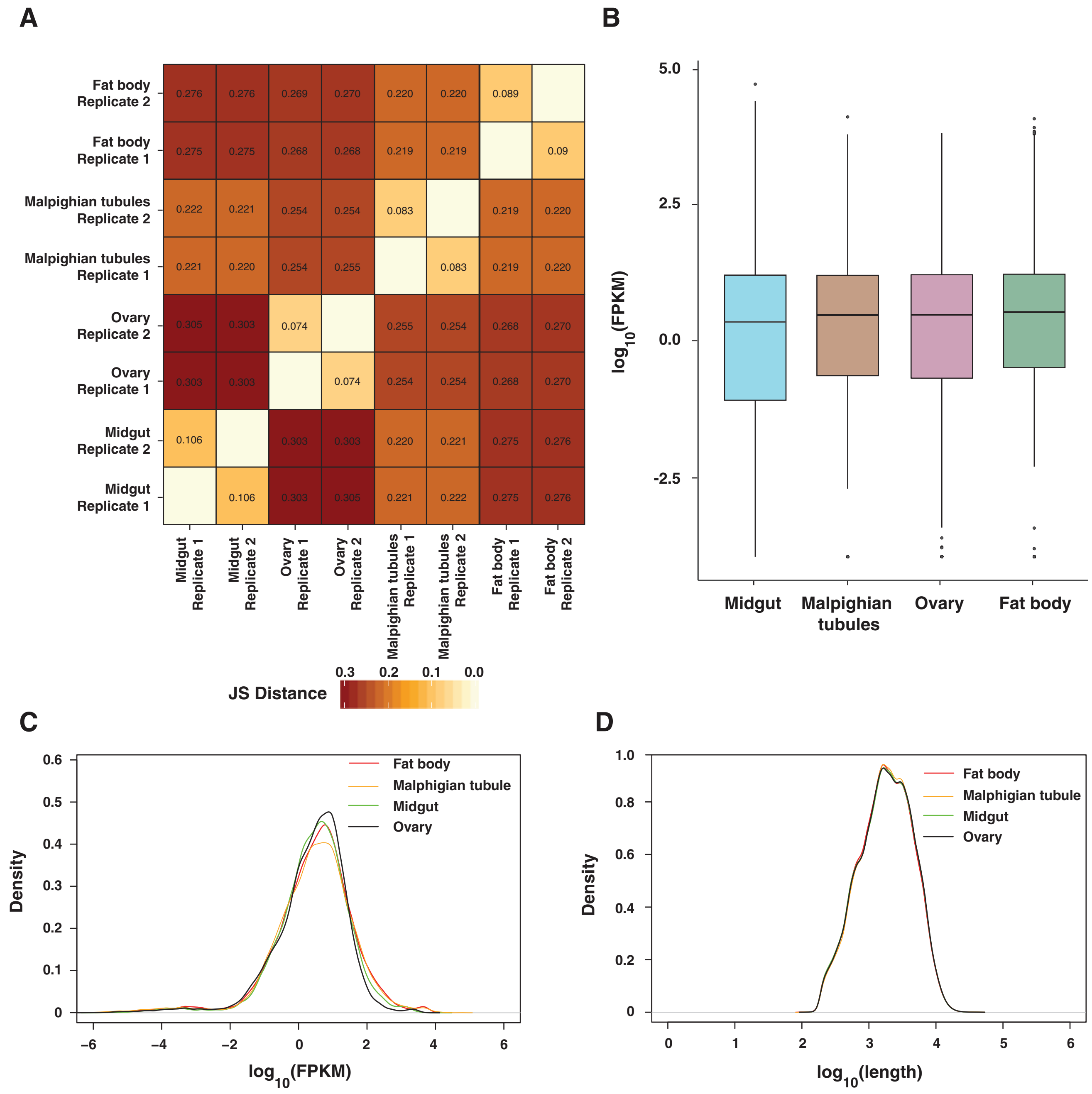

Fat body
Midgut
Malpighian tubules
Ovary
RNA isolation and indexed library preparation
NextGen Sequencing in replicates
Read alignment using HISAT2
Transcript assembly using StringTie
Analyze differential expression across tissues
log10 FPKM
log10 FPKM
Expression patterns and Transcript clusters

FIG. 1. The workflow representation of the study pipeline.



	Midgut Replicate 1	Midgut Replicate 2	Ovary Replicate 1	Ovary Replicate 2	Malpighian tubules Replicate 1	Malpighian tubules Replicate 2	Fat body Replicate 1	Fat body Replicate 2
Fat body Replicate 2	0.276	0.276	0.269	0.270	0.220	0.220	0.089	
Fat body Replicate 1	0.275	0.275	0.268	0.268	0.219	0.219		0.09
Malpighian tubules Replicate 2	0.222	0.221	0.254	0.254	0.083		0.219	0.220
Malpighian tubules Replicate 1	0.221	0.220	0.254	0.255		0.083	0.219	0.220
Ovary Replicate 2	0.305	0.303	0.074		0.255	0.254	0.268	0.270
Ovary Replicate 1	0.303	0.303		0.074	0.254	0.254	0.268	0.270
Midgut Replicate 2	0.106		0.303	0.303	0.220	0.221	0.275	0.276
Midgut Replicate 1		0.106	0.303	0.305	0.221	0.222	0.275	0.276

FIG. 2. The overall representation of transcript expression. **(A)** Heatmap representation of the Jensen–Shannon (JS) divergence between the different tissues and their technical replicates. **(B)** Bar-chart representation of the tissue-based transcripts and their median expression in the $\log_{10}$ (FPKM), showing normalized distribution. **(C)** FPKM distribution curve of the transcripts identified in the four tissues. **(D)** Distribution of transcript length across the four tissues. FPKM, Fragments per Kilobase of exon per Million Fragments Mapped.

to follow a tissue restricted expression pattern. Figure 3 details the distribution of the transcript expression (expressed with FPKM values ≥0.1) among the previously annotated transcripts (Fig. 3A), alternative isoforms (Fig. 3B), and novel previously unannotated intergenic transcripts (Fig. 3C). Among the known/annotated transcripts identified, 241 were found to be exclusive to midgut, 221 to Malpighian tubules, 479 transcripts to ovary, and 436 to fat body. The distribution of tissue-specific transcripts was also found to be similar among the alternative isoforms and novel intergenic transcripts of these four tissues, with 61, 68, 146, and 77 isoforms exclusively identified in midgut, Malpighian tubules, ovary, and fat body, respectively.

In general, there was a clear bias in the number of transcripts and transcript isoforms that were found to be common between midgut and Malpighian tubules and similarly between fat body and ovary than among the others. The diversity of the transcripts identified was found to be maximum in ovary, with most of the transcripts being identified in this tissue, followed by fat body. Midgut had the minimal number

Table 1. Transcript Distributions, the Number of Transcripts (Fragments per Kilobase of Exon per Million Fragments Mapped ≥0.1) in Total, and Class Code-Based Classification of Transcripts in All Four Tissues Shown Individually

	All four tissues	*Midgut*	*Malpighian tubules*	*Ovary*	*Fat body*
Total number of transcripts identified	23,009	17,461	18,812	18,616	18,685
Corresponding gene location identified	12,256	10,357	11,107	10,973	11,371
Total number of known/annotated transcripts—‘‘=’’	9722	7508	7883	8001	8015
Number of alternate isoforms/transcripts—‘‘j’’	8820	7603	8232	7992	8037
Number of novel transcripts (intergenic)—‘‘u’’	2694	2136	2458	2396	2398

of transcripts identified; however, the expression levels of these transcripts, in terms of FPKM, were higher than that of other tissues.

Novel splice variants and their expression

Apart from the known/annotated transcripts, we identified many splice (exon–exon) junctions that were not previously annotated. Assembly of exon spanning reads along with the intraexonic reads led to the identification of 8,820 transcripts that were spliced differently. These alternatively spliced isoforms represent the complexity of the transcript forms and their expression among all the four tissues. A summary of the differential expression of these alternate isoforms is provided in Figure 3B. As in the case of annotated transcripts, most of the alternatively spliced forms were also expressed in all the four tissues. Only about 1–2% of the total alternately spliced isoforms were found to have tissue-restricted expression. Transcript isoforms were enriched maximally in ovaries compared to any other tissue. With 146 isoforms restricted to ovaries, it showed the highest variation in the spliced forms among the four tissues although the FPKM values for these were comparatively lower than that of other tissues. Fat body had the least representation of the alternate isoforms.

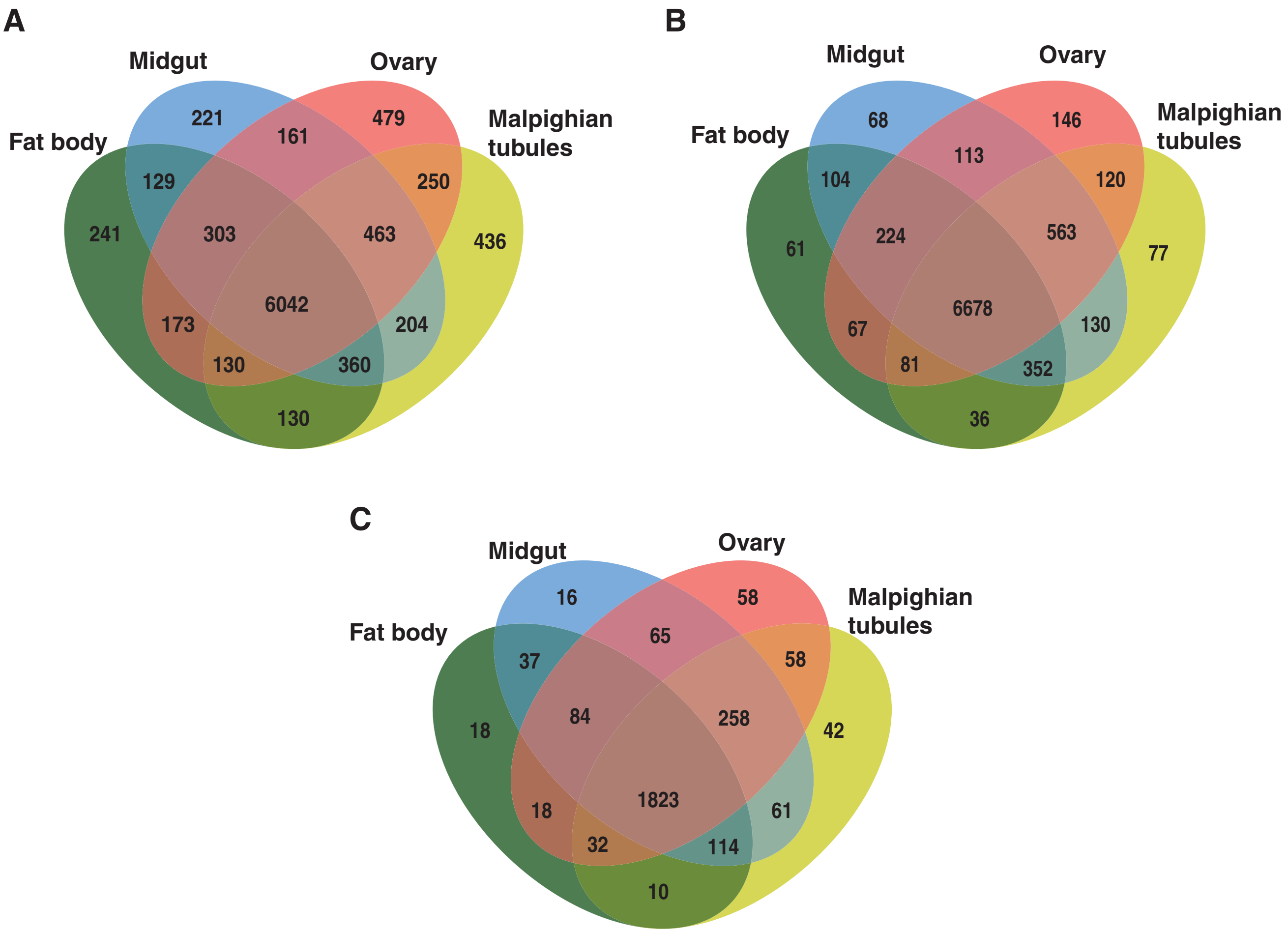


FIG. 3. Venn diagram representation depicting the overlap and tissue-specific expression of transcripts across the four tissues. **(A)** For VectorBase annotated transcripts. **(B)** Distribution of alternate isoforms of transcripts. **(C)** Distribution of novel intergenic transcripts.

The splice variants identified included examples of intron retention, alternative 3′ or 5′ donor and acceptor sites, exon skipping, and others. Different spliced forms were expressed in different tissues. An example of transcript expressed in different tissues is provided in Figure 4. The annotated gene ASTEI04270 belongs to the gelsolin/vilin/fragmin superfamily, coding for a single transcript isoform according to the VectorBase annotation.

However, we identified six different isoforms for this gene. The original protein, coded by the annotated transcript, is with a signal peptide and nine gelsolin-like domains and was found to be highly expressed in fat body, followed by Malpighian tubules. The alternative isoforms included a shorter transcript encoded by the first three exons (ANSTF.3986.3), which retained only three of the nine gelsolin-like domains along with the signal peptide sequence. It was found to be highly expressed in fat body with least expression in ovaries. The other four isoforms encoding the exons from fourth exon consist of four gelsolin-like domains.

Isoform ANSTF.3986.1 was found to be highly expressed in Malpighian tubules, followed by midgut. ANSTF.3986.2 was found to be highly expressed in Midgut, followed by Malpighian tubules. Both the isoforms had low expression in fat body and were not detected at all in ovaries. Two other isoforms ANSTF. 3986.4 and ANSTF.3986.5 were significantly expressed only in midgut. Proteins encoded by this superfamily typically consist of three to six gelsolin-like domains, with each domain playing a critical role in actin filament remodeling (Qian et al., 2015; Sun et al., 1999). Vast variations observed in the differential expression of many forms of alternate transcripts for a large number of genes will lead to new discoveries in the molecular complexity involved in the complex biological traits such as vectorial capacity and host–vector–pathogen interaction.

Novel intergenic transcripts

In addition to annotated and alternate spliced forms of the transcripts in the known/annotated gene loci, we found additional loci in the genome of *An. stephensi* Indian strain. The reads mapping to these unannotated regions were processed to assemble putative transcripts that were categorized as novel/unannotated transcripts. We identified 2,700 transcripts with FPKM values above 0.1 in the intergenic regions of the genome that were previously considered to be nontranscribed. The expression of most of these intergenic transcripts was found to be similar in all the four tissues. However, <1% of such transcripts were shown to have a tissue-restricted expression. The distribution of the identified transcripts was found to be similar to that of the annotated transcripts and isoforms, more intergenic transcripts were identified in ovary

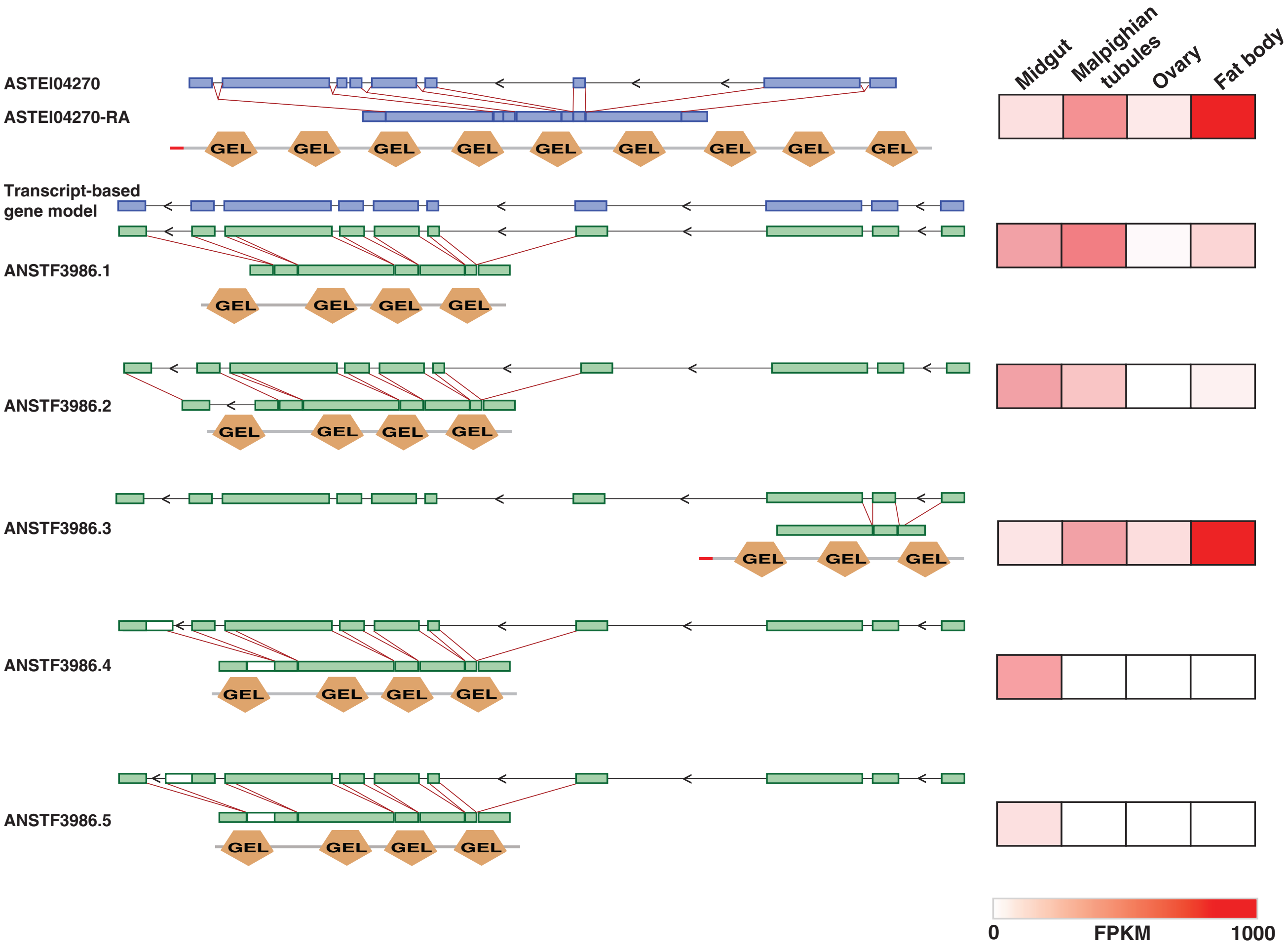


FIG. 4. An example representing the novel spliced forms of the VectorBase annotated gene ASTEI04270. Isoforms identified due to various splicing events and their expression across the four tissues.

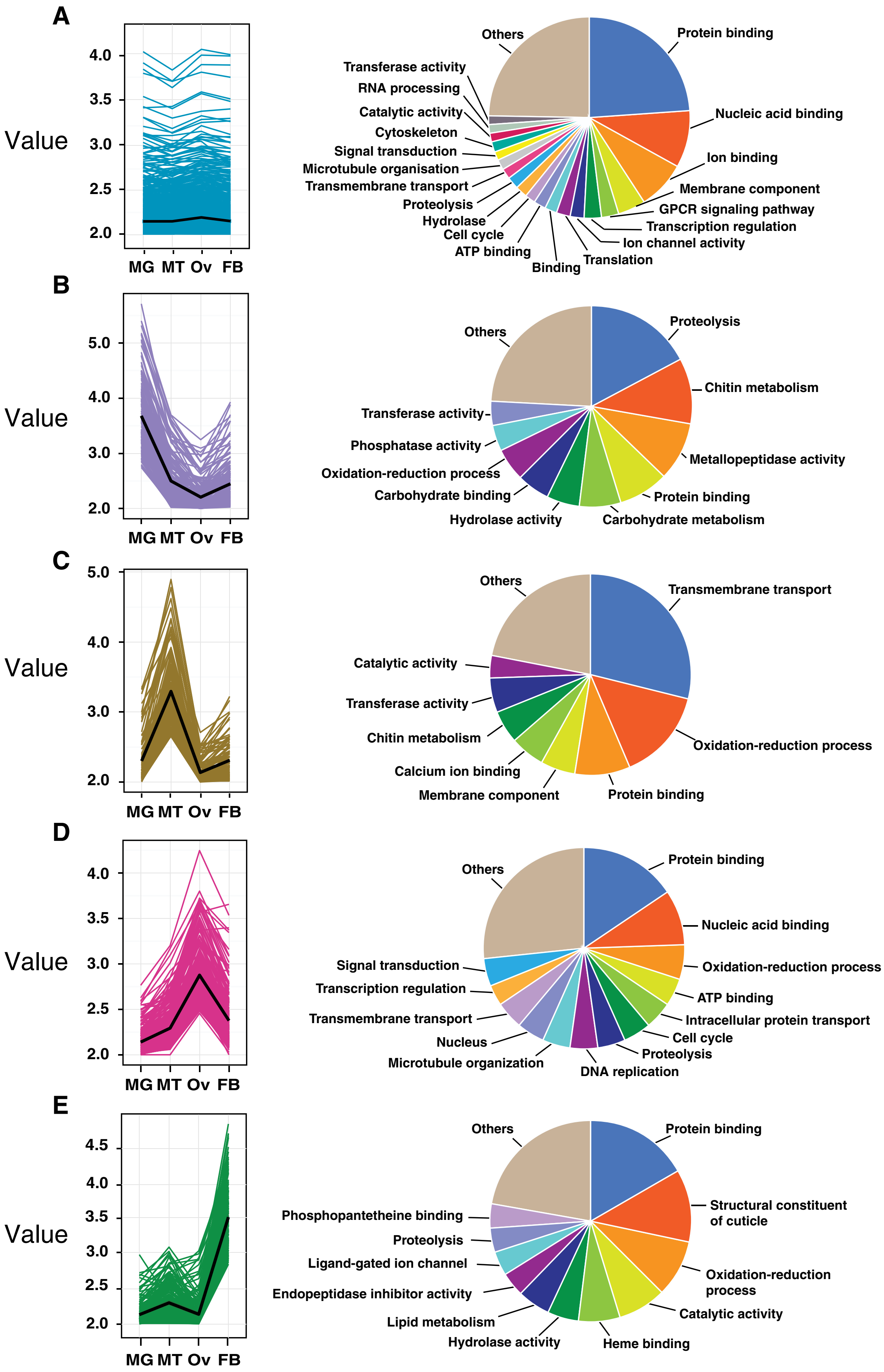


FIG. 5. Expression-based transcript clusters and the functional enrichment of the classes of transcripts based on domain and Gene Ontology-based functional annotation. **(A)** Transcripts having similar expression in all four tissues. **(B)** Midgut-enriched transcripts. **(C)** Transcripts overexpressed in Malpighian tubules. **(D)** Transcripts highly expressed in Ovary. **(E)** Fat body-enriched transcripts.

followed by fat body while majority of them (~1800) were common to all four tissues.

Identification and expression of lncRNAs

We compared the list of transcripts identified in our study with the list of transcripts that are annotated as noncoding RNAs (ncRNAs) in VectorBase. However, we failed to identify any of the annotated ncRNAs in our study since the annotated ones are largely ribosomal RNAs (rRNAs) and other small ncRNAs.

Due to the ribosomal RNA depletion used in our study, we expected no rRNAs to be identified. However, to investigate the presence and expression of the lncRNAs in our dataset, we assessed the coding potential of all the identified transcripts using the CPAT. From this, we identified 4,071 transcripts that satisfied the criteria for the lncRNAs (Supplementary Table S3). That is, they were longer than 200 bases in length and were predicted to have a coding potential of <0.39, which is used as the threshold for the lncRNAs according to the prebuilt models of classification for flies in CPAT. FPKM-based expression analysis of these transcripts in the four tissues showed that there was no significant difference between the expression levels of coding and noncoding transcripts (Supplementary Fig. S2).

Discussion

An. stephensi genome has been recently sequenced. It is relatively understudied in terms of postgenomic analyses and applications toward global health innovation. In particular, there is limited information on the function of these genes and predicted transcripts. Lack of experimental evidence for these transcripts and spliced forms have limited our knowledge to the current reference annotation, which is inferred from computational prediction algorithms. We tried to bridge this gap by presenting the RNA-seq-based evidence for over 23,000 transcripts and their differential expression in four tissues.

To understand the putative functional role of these transcripts, we examined *Gene Ontology* information available from VectorBase and domain prediction (InterProScan) of translated proteins. The transcripts were grouped into different clusters based on their expression pattern in the four tissues. Figure 5A represents a cluster of transcripts consisting of about 950 transcripts with similar expression pattern in all the four mosquito tissues. Gene level ontology mapping of these transcripts showed that majority of the transcripts from this cluster possessed generic domains such as protein, nucleotide and ion binding domains, transmembrane, transport, proteolysis, oxidoreductase activity, and signal transduction (Fig. 5A).

However, transcripts that were relatively overexpressed in each tissue represent the biological significance of the tissue in the vector life cycle. For example—genes which are involved in proteolysis and have peptidase activity were found to be enriched in midgut and are known to be involved in digestion of blood meal. Host-derived proteins present in the blood meal are digested by the peptidase activity of such proteins, which makes free amino acids available for egg development. In addition, various proteases and metallopeptidases are known to affect *Plasmodium* life cycle in the vector (Goulielmaki et al., 2014; Jahan et al., 1999; Sriwichai et al., 2012). Some of the midgut-enriched transcripts were found to be involved in chitin and carbohydrate metabolism (Fig. 5B).

Similarly, 116 transcripts that were enriched in Malpighian tubules were found to be largely associated with transmembrane transportation, oxidation-reduction process, and protein- and ion-binding events. Several transmembrane proteins known as septate junctional proteins are involved in detoxification process and have been shown to differentially express upon blood meal digestion (Esquivel et al., 2016). Some of the transcripts were also associated with transferase, ligase, and lyase activities among other catalytic activities (Fig. 5C). Ovary-enriched transcript cluster of 241 were found to be associated with the protein binding, nucleic acid, and ATP binding. In addition, these transcripts were also found to have signaling and transport domains associated with intracellular signal transduction processes such as G-protein-coupled receptor (GPCR) activity, protein phosphorylation, and dimerization.

As expected, these transcripts seem to be involved highly in cell cycle processes, including DNA replication, microtubule organization, DNA repair, and growth factor activities, which are crucial mechanisms for egg development (Fig. 5D). Fat body-enriched transcripts (170) were consistent with the role of fat body akin to the vertebrate liver and found to be associated majorly with transmembrane transportation, oxidation-reduction process, chitin binding and metabolism, heme binding and transport, in addition to oxidoreductase activities (Fig. 5E).

Tissues considered in this study play a significant role in the life cycle of the female mosquito. They are found to be critical in the digestion of blood meal, metabolism, vitellogenesis, excretion, immunogenesis, *Plasmodium* sporogony, and reproduction, which are known to be associated with vector physiology, progression, and malaria transmission. Mosquito midgut is known to be involved in the initial storage and digestion of the ingested blood. The gut epithelium also provides as site for development of oocysts and sporozoites (sporogony). Blood meal is known to induce pathways such as TOR, which ultimately leads to synthesis of proteins required for egg development. Fat body and ovary are known to be involved in the utilization of the nutrients from blood to enable vitellogenesis. Malpighian tubules along with the hindgut form the excretory system of the mosquitoes and are known to be rich in the solute transporters (Dow, 2009; Piermarini et al., 2017).

Likewise, in our study, we noticed that the transcripts that were enriched in Malpighian tubules belonged to the class of transmembrane solute transporters and detoxification genes, including cytochrome p450 class of genes (Fig. 5C). Fat body cells (trophoblasts) and Malpighian tubules have also been shown to be involved in the immune responses (Dow, 2009; Martins et al., 2011; Piermarini et al., 2017; Verma and Tapadia, 2012). Therefore, these organs are now being considered as targets for mosquito control (Dow, 2009; Piermarini et al., 2017). Toward this end, we further evaluated the expression of *An. stephensi* orthologs of *An. gambiae* genes that were previously reported to be involved in the vector–pathogen interactions (Sreenivasamurthy et al., 2013) across the four tissues (Supplementary Table S4).

Conclusions and Outlook

The affordability and accessibility of sequencing-based techniques have resulted in numerous transcriptome-based studies even in *An. stephensi* (Biedler et al., 2014; Jiang et al., 2014, 2017; Thomas et al., 2016). However, majority of the studies have been performed on whole mosquitoes (Jiang et al., 2014, 2017; Neafsey et al., 2015), which can act only as a proof of expression of predicted transcripts from the genome and do not provide biological insights into their function. Tissue level transcript expression studies have been studied in salivary glands (Dixit et al., 2011) using complementary DNA (cDNA) libraries; by RNA-seq in ovaries (Biedler et al., 2014; Jiang et al., 2014) and hemocytes (Thomas et al., 2016) of *An. stephensi*.

However, due to the low depth of RNA-seq data in these studies, no significant comparison could be performed with data from the current study. Several novel splice variants reported in this study were previously not described and could have been easily missed due to lack of tissue-wise transcriptomic profiling approach. Further investigation on the transcripts that were found to be enriched in each tissue will provide a novel insight into the functional contribution of these tissues in vector life cycle and malaria transmission. A subset of transcripts identified in our study is found to be unique to *An. stephensi* and does not have orthologs in other vector species. Such information will provide a platform for future studies for selective and targeted control of *An. stephensi* population.

Novel intergenic transcripts identified in our study will add to the existing knowledge of genes encoded by *An. stephensi* genome and their selective expression in different tissues. Considering such information, analysis of gene expression data in the context of changes due to blood meal and infection might lead to new perspectives and insights in vector–pathogen interaction and disease transmission. This, in turn, will facilitate the choice of novel targets for vector control and transmission blocking studies and other experiments as evidenced in *An. gambiae* (Domingos et al., 2017).

Data availability

The RNA-sequencing data have been submitted to the Sequence Read Archive (SRA) from NCBI and can be accessed using the project accession number SRP043489. The processed files have been uploaded to Gene Expression Omnibus (GEO) and can be accessed using the accession number GSE99679.

Acknowledgments

This study is funded by the joint research project to NIMR and Institute of Bioinformatics (IOB) entitled ‘‘Characterization of Malaria Vector *Anopheles stephensi* Proteome and Transcriptome’’ (EMR/2014/000444) from the Science and Engineering Research Board (SERB), Government of India. The funders had no role in the design, implementation, analysis, and/or reporting of the present work. S.K.S. and G.D. have been supported by the Senior Research Fellowship by University Grants Commission (UGC), A.K.M. was supported by the Department of Biotechnology, and M.K. was supported by the Council of Scientific and Industrial Research, Government of India during the study.

Author Disclosure Statement

The authors declare that no conflicting financial interests exist.

References

Biedler JK, Qi Y, Pledger D, James AA, and Tu Z. (2014). Maternal germline-specific genes in the Asian malaria mosquito *Anopheles stephensi*: Characterization and application for disease control. G3 (Bethesda) 5, 157–166.

Chaerkady R, Kelkar DS, Muthusamy B, et al. (2011). A proteogenomic analysis of *Anopheles gambiae* using high-resolution Fourier transform mass spectrometry. Genome Res 21, 1872–1881.

Childs LM, Cai FY, Kakani EG, et al. (2016). Disrupting mosquito reproduction and parasite development for malaria control. PLoS Pathog 12, e1006060.

Dixit R, Rawat M, Kumar S, Pandey KC, Adak T, and Sharma A. (2011). Salivary gland transcriptome analysis in response to sugar feeding in malaria vector *Anopheles stephensi*. J Insect Physiol 57, 1399–1406.

Domingos A, Pinheiro-Silva R, Couto J, do Rosario V, and de la Fuente J. (2017). The *Anopheles gambiae* transcriptome—A turning point for malaria control. Insect Mol Biol 26, 140–151.

Dow JA. (2009). Insights into the Malpighian tubule from functional genomics. J Exp Biol 212, 435–445.

Esquivel CJ, Cassone BJ, and Piermarini PM. (2016). A de novo transcriptome of the Malpighian tubules in non-blood-fed and blood-fed Asian tiger mosquitoes *Aedes albopictus*: Insights into diuresis, detoxification, and blood meal processing. Peer J 4, e1784.

Gantz VM, Jasinskiene N, Tatarenkova O, et al. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. Proc Natl Acad Sci U S A 112, E6736–E6743.

Goff L, Trapnell C, and Kelley D. (2013). cummeRbund: Analysis, exploration, manipulation, and visualization of Cufflinks high-throughput sequencing data. R package version 2.18.0 (www.bioconductor.org).

Goulielmaki E, Sidén-Kiamos I, and Loukeris TG. (2014). Functional characterization of Anopheles matrix metalloprotease 1 reveals its agonistic role during sporogonic development of malaria parasites. Infect Immun 82, 4865–4877.

Jahan N, Docherty PT, Billingsley PF, and Hurd H. (1999). Blood digestion in the mosquito, *Anopheles stephensi*: The effects of *Plasmodium yoelii* nigeriensis on midgut enzyme activities. Parasitology 119, 535–541.

Jiang X, Hall AB, Biedler JK, and Tu Z. (2017). Single molecule RNA sequencing uncovers trans-splicing and improves annotations in *Anopheles stephensi*. Insect Mol Biol 26, 298–307.

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

Kelkar DS, Provost E, Chaerkady R, et al. (2014). Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis. Mol Cell Proteomics 13, 3184–3198.

Kim D, Langmead B, and Salzberg SL. (2015). HISAT: A fast spliced aligner with low memory requirements. Nat Methods 12, 357–360.

Kumar A, Hosmani R, Jadhav S, et al. (2016). *Anopheles subpictus* carry human malaria parasites in an urban area of Western India and may facilitate perennial malaria transmission. Malar J 15, 124.

Martins GF, Serrao JE, Ramalho-Ortigao JM, and Pimenta PF. (2011). A comparative study of fat body morphology in five mosquito species. Mem Inst Oswaldo Cruz 106, 742–747.

Mitchell SN, Kakani EG, South A, Howell PI, Waterhouse RM, and Catteruccia F. (2015). Mosquito biology. Evolution of sexual traits influencing vectorial capacity in anopheline mosquitoes. Science 347, 985–988.

Neafsey DE, Waterhouse RM, Abai MR, et al. (2015). Mosquito genomics. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. Science 347, 1258522.

Pertea M, Kim D, Pertea GM, Leek JT, and Salzberg SL. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc 11, 1650–1667.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, and Salzberg SL. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33, 290–295.

Piermarini PM, Esquivel CJ, and Denton JS. (2017). Malpighian tubules as novel targets for mosquito control. Int J Environ Res Public Health 14, E111.

Prasad TS, Mohanty AK, Kumar M, et al. (2017). Integrating transcriptomic and proteomic data for accurate assembly and annotation of genomes. Genome Res 27, 133–144.

Qian D, Nan Q, Yang Y, et al. (2015). Gelsolin-like domain 3 plays vital roles in regulating the activities of the lily villin/gelsolin/fragmin superfamily. PLoS One 10, e0143174.

Ramasamy R, and Surendran SN. (2016). Mosquito vectors developing in atypical anthropogenic habitats: Global overview of recent observations, mechanisms and impact on disease transmission. J Vector Borne Dis 53, 91–98.

Sinka ME, Bangs MJ, Manguin S, et al. (2012). A global map of dominant malaria vectors. Parasit Vectors 5, 69.

Sougoufara S, Doucouré S, Backé Sembéne PM, Harry M, and Sokhna C. (2017). Challenges for malaria vector control in sub-Saharan Africa: Resistance and behavioral adaptations in *Anopheles* populations. J Vector Borne Dis 54, 4–15.

Sreenivasamurthy SK, Dey G, Ramu M, et al. (2013). A compendium of molecules involved in vector-pathogen interactions pertaining to malaria. Malar J. 12, 216.

Sriwichai P, Rongsiryam Y, Jariyapan N, et al. (2012) Cloning of a trypsin-like serine protease and expression patterns during *Plasmodium falciparum* invasion in the mosquito, *Anopheles dirus* (Peyton and Harrison). Arch Insect Biochem Physiol 80, 151–165.

Sun HQ, Yamamoto M, Mejillano M, and Yin HL. (1999). Gelsolin, a multifunctional actin regulatory protein. J Biol Chem 274, 33179–33182.

Thomas T, De TD, Sharma P, et al. (2016). Hemocytome: Deep sequencing analysis of mosquito blood cells in Indian malarial vector *Anopheles stephensi*. Gene 585, 177–190.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, and Pachter L. (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 31, 46–53.

Venkatrao V, Kumar SK, Sridevi P, Muley VY, and Chaitanya RK. (2017). Cloning, characterization and transmission blocking potential of midgut carboxypeptidase A in *Anopheles stephensi*. Acta Trop 168, 21–28.

Verma P, and Tapadia MG. (2012). Immune response and antimicrobial peptides expression in Malpighian tubules of *Drosophila melanogaster* is under developmental regulation. PLoS One 7, e40714.

Wang L, Park HJ, Dasari S, Wang S, Kocher JP, and Li W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. Nucleic Acids Res 41, e74.

World Health Organization. World Malaria Report, 2016. http://apps.who.int/iris/bitstream/10665/252038/1/9789241511711-eng.pdf Accessed March 7, 2017.

Address correspondence to:
Thottethodi Subrahmanya Keshava Prasad, PhD
Institute of Bioinformatics
Discoverer Building, 7th Floor
International Tech Park
Whitefield
Bangalore 560 066
Karnataka
India

E-mail: keshav@ibioinformatics.org

Abbreviations Used

BAM	= Binary Alignment Map
CPAT	= Coding Potential Assessment Tool
FPKM	= Fragments per Kilobase of exon per Million Fragments Mapped
GTF	= Gene Transfer File
lncRNAs	= long non-coding RNAs
ncRNA	= non-coding RNA
NIMR	= National Institute of Malaria Research
rRNA	= ribosomal RNA
UTR	= untranslated region