Anopheles stephensi tissue-restricted expression

Tissue-restricted expression and alternative splicing revealed by transcriptome profiling of *Anopheles stephensi*

- Sreelakshmi K. Sreenivasamurthy [1, 2] , Anil Madugundu [1, 3] , Arun Kumar Patil [1, 4, 5] , Gourav Dey [1, 2] , Ajeet Kumar Mohanty [6] , Manish Kumar [1, 2] , Krishna Patel [1] , Charles Wang [7] , Ashwani Kumar [6] , Akhilesh Pandey [1, 8, 9, 10, 11] , T. S. Keshava Prasad [1, 2, 4,*]

Abstract

The sequencing of *Anopheles stephensi* , a major malaria vector in Asia has led to increased research activity to understand the vectorial ability of this mosquito species. However, tissue-based gene expression profiles of the annotated genes remain to be understood. In this study, we summarize the transcriptomic profile of four important organs of a female imago – Midgut, Malpighian tubules, Fat body and Ovary. We identified over 21, 000 transcripts in total, from all the four tissues corresponding to about 12, 000 gene loci. This study provides an account of the tissue-based expression profiles of majority of annotated transcripts in *An. stephensi* genome and alternative splicing in these tissues. Understanding of the transcript expression and gene function at the tissue level would immensely help in enhancing our knowledge of this important vector and decipher the putative role of these mosquito tissues, providing the basis of selection of candidates for future studies on vectorial ability.

Keywords: Mosquito, RNA-seq, differential expression, lncRNAs

Introduction

Malaria remains as one of the most debilitating mosquito-borne diseases till date. According to WHO World Malaria Report in 2016, there were ~212 million malaria cases in the year 2015, resulting in an estimated death of about 429, 000 individuals globally. Most of these cases (90%) is in the African region with about 7% incidence in South East Asia. About 50% of the Asian malaria incidence and deaths has been in India [1] . The number of deaths attributed to malaria in India has been reported to be higher than the WHO estimates [2] . However, the latest updates on the cases and deaths reported in India has been limited to the National Vector Borne Disease Control Programme (NVBDCP), according to which there has been about a million cases of malaria reported in the year 2014 [http://www. nvbdcp. gov. in/malaria3. html]. Out of the 41 different *Anopheline* species reported as significant vectors for transmission of human malaria, *An. stephensi* is an important vector in India and South Asia [3, 4] . Being the major urban vector, it is second most prevalent in India. It has been reported all over the country except the north-eastern states of Sikkim, Arunachal Pradesh, Mizoram, Nagaland, Manipur and Tripura [5] .

Sequencing of the Anopheles mosquito genomes have resulted in a spurt of activity in the study of Anopheles mosquitoes. PubMed search with the keyword “ Anopheles” resulted in 14, 576 publications, majority of which have been after the year 2000 as shown in the Figure 1A. Majority of the studies post-genome sequencing has been focussed towards understanding the role of various genes and development of numerous methods to regulate their expression. The overall aim of the community is to embark on a

feasible means to control the spread of infectious organisms either by controlling the vector/mosquito population or by curbing or reducing their vectorial ability. In this regards, numerous studies have already been performed on the recently sequenced malarial vectors [6-9] . However, most of the studies are focussed on previously studied molecules with very few studies focussing on new target molecules. This could probably be due to the lack of reliable data owing to incomplete genome assemblies and annotations in the identification of such targets. We have tried to bridge this gap with a huge effort of supplementing the current efforts using an integrated approach of utilizing proteomic and transcriptomic data in the genome annotation and assembly in an array of organisms through our previous studies [10-12] . Although transcriptomic data played a major role in refining the annotations and assembly of the genomes in the previous study, the tissue-based expression profiles were not focussed on. The tissue-based expression profiles of the identified transcripts are the focus of this study.

Tissue-based expression profiling in *An. stephensi* has been limited to salivary glands [13] , ovaries [14, 15] , testes [16] and hemocytes [17] with most the studies being done on whole mosquitoes [14, 18, 19] . However, even with the availability of transcriptome data from ovaries, there are several issues. The study was mainly focused on identification of transcripts expressed in developmental stages and is of low throughput [14] . The other tissue-based expression studies published including one cDNA-based study of the salivary glands and another being cDNA sequencing of the transcripts from hemocytes, both tissues that were not included in our study. The focus of our study is on the Midgut, ovary, Malpighian tubule and fat body of a female *An.*

stephensi imago. These tissues, along with the salivary glands play a very important role in the blood meal digestion and thus important for the life cycle of the mosquito vector and plasmodium species. However, we restrict ourselves to understand the molecular difference between these mosquito tissues in the uninfected sugar-fed state of the mosquito which we believe will provide the much-needed basic understanding of the role played by these tissues. To this extent, we performed comparative and deep transcriptomic analysis of these four tissues.

Materials and Methods

RNA isolation and sequencing

Adult female *An. stephensi* mosquitoes grown at the NIMR field station, Goa, were dissected to obtain midgut, Malpighian tubules, ovaries and fat body. These dissected tissues were stored in RNAlater to preserve the RNA quality till RNA extraction. The RNA isolation and sequencing was performed as described earlier [10, 11] . Briefly, the RNA isolated using Qiagen miRNeasy kit was used for the preparation of indexed RNA-seq libraries using TruSeq RNA Sample Preparation Kit v3. The indexed and pooled libraries were sequenced on two lanes (as technical replicates) of Illumina HiScan SQ platform.

Read alignment and transcript assembly

The raw reads were processed for quality filtration 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. FastQC (Version 0. 10. 1) tool was used to determine the quality of the raw data and poor quality calls with Phred score <20 was trimmed off using fqtrim v0. 9. 4. Post-trimming, reads that were less than 60 bp were discarded to avoid ambiguous alignment.

Quality filtered reads were aligned against *A n. stephensi* genome build (ASTEI2) downloaded from VectorBase (https://www. vectorbase. org/) using HISAT (Version 2. 1. 0) [20] 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 [21] . Assembled transcripts were further quantified and annotated into known and novel categories using the ‘ gffcompare’ in StringTie package as described earlier [22] . To determine novel transcripts as a transcript GTF file and all the StringTie assemblies 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 by the use of the Coding Potential Assessment Tool (CPAT) [23] . Transcripts which were > 200 bp in length with a CPAT score threshold of <0. 39 was categorized as long non-coding RNAs according to the specifications provided for the fly database.

Identification of differentially expressed genes across four tissues

Merged GTF file from StringTie was annotated in to 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 were 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) [24] . The R-package version 2. 16. 0 of cummeRbund was used for visualization, analysis of RNA-seq data and cluster generation [25] . An overview of the analysis pipeline is provided in Figure 1B. To identify tissue specific transcripts, we initially filtered transcripts with FPKM value â ‰¥ 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 relatively high in abundance in one tissue as against other tissues.

Results and Discussion

Transcriptome sequencing of four *An. stephensi* tissues – Midgut, Malpighian tubules, Fat body and Ovary was performed to create a tissue-based expression profile. In total, about 500 million paired-end reads of 100bp were generated from all the four tissues, with about 55 million read pairs per tissue sample from two lanes. The expression levels of transcripts between the replicates and among the tissues were comparable. Figure 2A represents the inter-tissue and intra tissue transcript expression variations in the form of a distance-based heatmap. The variations are minimal between the replicates as expected and increases between the tissues with Ovary and Malpighian tubules being the most different. By following the standard alignment and assembly pipeline using the HiSAT2 and StringTie assembler, we identified a total of about 25, 000 transcripts. However, after the initial filtering for the FPKM values (â‰¥ 0. 1) only 21, 500 transcripts were

retained. The expression of these transcripts was comparable across tissues with the median FPKM value ranging about 2 to 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 transcript assemblies were found to have FPKM value of 1 and above, while the average length of majority of the transcripts tend to be in the range of 1000 to 3000 bp. This shows an expected trend of a reliable depth and absence of any skewing. The Transcript assemblies were classified into different classes using gffcompare. However, in order to avoid over interpretation 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.

In our analysis, we noticed that almost equivalent number of transcript assemblies were classified under the known (=) and the alternate (j) categories. In fact, the transcript assemblies in the “ j” category exceeded the number of known transcript assemblies. A deeper look in to this matter showed us that due to the poorly annotated gene models (which is mostly based on the prediction program) for this strain, the untranslated regions (UTRs) of the predicted transcript models in the current annotation is missed. As a result, the transcript assemblies with the extension of the exonic regions supported by the reads, probably into the UTRs were classified as alternate transcripts. We are working closely with the VectorBase to improve the annotations of these predicted gene and transcript models for the *An. stephensi* Indian strain.

Tissue restricted transcripts

Majority of the transcripts identified (about 87%) were expressed largely at similar levels in all the four tissues, the remaining 15% of the transcripts identified seemed to have more of a tissue restricted expression. Figure 3 details the distribution of the transcript expression (expressed with FPKM values â‰¥ 0. 1) among the previously annotated transcripts (Figure 3A), alternative isoforms (Figure 3B) and novel previously unannotated intergenic transcripts (Figure 3C). The majority of the transcripts in each of these groups are expressed in all the four tissues with only about 3 – 4% of the transcripts showing tissue restricted expression. Among the known/annotated transcripts identified, 241 were found to be exclusive to Midgut, 221 exclusive to Malpighian tubules, 479 transcripts in Ovary and 436 in Fat body. The distribution of tissue specific transcripts was similar in the alternative isoforms and novel intergenic transcripts of these four tissues with 61, 67, 146 and 77 isoforms exclusively identified in Midgut, Malpighian tubules, Ovary and Fat body. In general, there was a clear bias in the number of transcripts and transcript isoforms that were common between midgut and Malpighian tubules and similarly between fat body and ovary than amongst the others. The diversity of the transcripts identified was found to be maximal in Ovary with most the transcripts being identified in this tissue, followed by fat body. Midgut had the minimal number 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 a plethora of spliced (exon-exon) reads that were not previously annotated. Assembly of

such reads along with the intra exonic reads led to the identification of > 8500 transcripts that were spliced differently. These alternatively spliced isoforms represent the complexity of the transcript forms and their expression in 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 alternate transcripts 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.

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 the gene. The original protein coded by the annotated transcript with a signal peptide and nine gelsolin-like domains that was highly expressed in Fat body followed by Malpighian tubules. The alternative isoforms included a shorter transcript encoded by the first 3 exons (ANSTF. 3986. 4), which retained only three of the nine gelsolin-like domains along with the signal peptide

sequence that was highly expressed in fat body and least expression in ovaries. The other 4 isoforms encoding the exons from fourth exon consists of 4 gelsolin-like domains. Isoforms ANSTF. 3986. 1 and ANSTF. 3986. 2 were highly expressed in midgut followed by Malpighian tubules but not identified in fat body and ovaries. Whereas, isoforms ANSTF. 3986. 5 and ANSTF. 3986. 6 were significantly expressed only in midgut. Proteins encoded by this superfamily typically consists of three to six gelsolin-like domains (GEL), with each domain playing a critical role in actin filament remodeling [26, 27] .

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 about 2700 transcripts with FPKM values above 0. 1 in the intergenic regions of the genome that were previously considered to be non-transcribed. The expression of most of these intergenic transcripts were found to be similar in all the four tissues. However, <1% of such transcripts were shown to have a tissue restricted expression (Figure 3C). 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 followed by Fat body while majority of them (about 1800) were common to all four tissues.

Expression-based clustering and functional correlation

Since *An. stephensi* genome was recently sequenced and is relatively less worked upon, there is limited information on the function of these genes and

transcripts. However, *Gene Ontology* analysis based on their translated protein and the domain structures (Interpro domains) showed that most of the differentially expressed transcripts were found to have expected domains as per the perceived function of these respective tissues.

The identified transcripts were segregated into clusters based on their expression levels in the four mosquito tissues. Among the various clusters generated using the cummerbund package, few of the clusters showed clear trends of expression. One of the clusters with about 950 transcripts showed similar expression in all the four tissues. Gene level ontology mapping of these transcripts showed that majority of the transcripts possessed generic domains such as protein, nucleotide and ion binding domains, transmembrane transport, proteolysis, oxidoreductase activity and signal transduction (Figure 5A). Transcripts found to be enriched in the Midgut (170) compared to other tissues were found to have proteolytic, protein binding, hydrolase and peptidase activity. Some of the midgut enriched transcripts were found to be involved in chitin and carbohydrate metabolism (Figure 5B). Transcripts enriched in Malpighian tubules (116) were found to be associated largely with transmembrane transportation, oxidation-reduction process, protein and ion binding events. Few of the transcripts were associated with transferase, ligase and lyase activities among other catalytic activities (Figure 5C). Ovary enriched transcripts (241) were associated more with the protein binding, nucleic acid and ATP binding, in addition to those having signaling domains and transport domains associated with intracellular signal transduction processes such as 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 vitellogenesis (Figure 5D). Fat body enriched transcripts (170) were consistent with the role of fat body akin to the vertebrate liver. The transcripts enriched in fat body are associated majorly with transmembrane transportation, oxidation-reduction process, chitin binding and metabolism, heme-binding and transport, in addition to oxidoreductase activities (Figure 5E).

Identification and expression of long non-coding RNAs

We compared the list of transcripts identified in our study to the list of transcripts that are annotated as non-coding RNAs in VectorBase. However, we failed to identify any of the annotated non-coding RNAs in our study since the annotated ones are largely rRNAs and other small ncRNAs. Due to the ribosomal RNA depletion employed in our study, we expected no rRNAs to be identified. However, in order to investigate the presence and expression of the long non-coding RNAs in An. stephensi, we assessed the coding potential of all the identified transcripts using the CPAT tool. From this, we identified 4, 071 transcripts that satisfied the criteria for the long non-coding RNAs (lncRNAs) (Supplementary Table 2). 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 in flies. FPKM-based expression analysis of these transcripts in the four tissues showed that there was no significant difference between the coding and non-coding transcript expression levels (Supplementary Figure 3).

Tissues considered in this study play an important role in the life cycle of the female mosquito. They are critical in blood meal digestion, metabolism, vitellogenesis, excretion, immunogenesis, Plasmodium sporogony and reproduction, which are associated with vector physiology, progression and malaria transmission. Mosquito midgut is involved in the initial storage and digestion of the ingested blood. The gut epithelium also provides site for development of oocysts and sporozoites (Sporogony). Blood meal induces 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 are known to play an important role in the mosquito xenobiotics. Fat body cells (trophoblasts) and recently, Malpighian tubules have also been shown to be involved in the immune responses [28-31] and is now being considered as targets for mosquito control [28, 31] . Towards this end, we further evaluated the expression of genes previously reported to be involved in the vector-pathogen interactions [32] across the four tissues (Table 2).

The affordability and accessibility of sequencing-based techniques have resulted in numerous transcriptome-based studies even in *An. stephensi* [14, 15, 17, 19] . However, due to the low depth of the other existing studies, no significant comparison could be performed between the transcript expression from our study to that of the other studies. We deciphered the genes reported to be involved in immunity [14] and evaluated the expression information for the annotated transcripts and the novel alternate isoforms

across the tissues (Supplementary Table 4). Although, there has been a recent study of the cDNAs from hemocytes, we could not compare the genes expressed in their study since hemocytes were not part of our study. Another reason for non-comparison was normalization issues caused by 36bp single end reads in their study, with only 49% of it mapping to the VectorBase assembly. We provide the deepest tissue-based transcriptome profiling for these four organs of *An. stephensi* (Indian strain), so far. Studies such as ours depicting the transcript variations amongst tissues in its physiological states provide important baseline information. In light of such information, analysis of gene expression data in the context of changes due to blood meal, infection of insecticide resistance might lead to new perspectives and insights. 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* [33] .

Data Availability

The RNA-sequencing data has been submitted to the Sequence Read Archive (SRA) from NCBI and can be accessed using the project accession number SRP043489.

Supplementary data

Supplementary data are available atwww. dnaresearch. oxfordjournal. org.

Funding

This paper is funded by the joint research project to NIMR and 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. SKS and GD has been

supported by the Senior Research Fellowship by University Grants Commission (UGC) and MK was supported by the Council of Scientific and Industrial Research, Government of India during the study.

Table 1. Transcript distribution – number of transcripts in total, class code-based classification of transcripts in all four tissues and in individual tissues

All 4 tissues	Midgut	Malpighian tubule	Ovary	Fat body	
Total number of transcripts identified	21, 500	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 – “=”	9, 722	7, 508	7, 883	8, 001	8, 0
Number of alternate isoforms/transcripts – “ j”	8, 820	7, 603	8, 232	7, 992	8, 0
Number of novel transcripts (intergenic) – “ u”	2, 694	2, 136	2, 458	2, 396	2, 3

Figure Legends:

Figure 1. A. Graphical representation of the remarkable increase in the number of studies on Anopheles mosquitoes post genomic era. B. Workflow representation of the study pipeline followed.

Figure 2. 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.

Figure 3. Venn diagram representation depicting the overlap and the tissue specific expression of the transcripts across the four tissuesA. For VectorBase annotated transcripts. B. Distribution of alternate isoforms of transcripts. C. Distribution of novel intergenic transcripts.

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

Figure 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 tissuesB. Midgut-enriched transcriptsC. Transcripts overexpressed in Malpighian tubulesD. Transcripts highly expressed in OvaryE. Fat body-enriched transcripts.