Transcriptome analysis of *Anopheles stephensi* embryo using expressed sequence tags

KAUSTUBH GOKHALE[1,†,‡], DEEPAK P PATIL[1,†,$], DHIRAJ P DHOTRE[1], RAJNIKANT DIXIT[1,§], MURLIDHAR J MENDKI[2], MILIND S PATOLE[1] and YOGESH S SHOUCHE[1,*]

[1]*National Centre for Cell Science, Ganeshkhind, Pune 411 007*

[2]*Entomology Division, Defense Research and Development Establishment, Gwalior 474 002*

[‡]*Present address: School of Life Sciences, Arizona State University, Tempe, AZ USA*

[$]*Present address: Center for RNA Biology, The Ohio State University, Columbus 43202 OH USA*

[§]*Present address: National Institute of Malaria Research, Sector 8, Dwarka, Delhi 110077, India*

**Corresponding author (Fax, +91-20-25692259; Email, yogesh@nccs.res.in)*

[†]*These authors contributed equally to the work.*

Germ band retraction (GBR) stage is one of the important stages during insect development. It is associated with an extensive epithelial morphogenesis and may also be pivotal in generation of morphological diversity in insects. Despite its importance, only a handful of studies report the transcriptome repertoire of this stage in insects. Here, we report generation, annotation and analysis of ESTs from the embryonic stage (16–22 h post fertilization) of laboratory-reared *Anopheles stephensi* mosquitoes. A total of 1002 contigs were obtained upon clustering of 1140 high-quality ESTs, which demonstrates an astonishingly low transcript redundancy (12.1%). Putative functions were assigned only to 213 contigs (21%), comprising mainly of transcripts encoding protein synthesis machinery. Approximately 78% of the transcripts remain uncharacterized, illustrating a lack of sequence information about the genes expressed in the embryonic stages of mosquitoes. This study highlights several novel transcripts, which apart from insect development, may significantly contribute to the essential biological complexity underlying insect viability in adverse environments. Nonetheless, the generated sequence information from this work provides a comprehensive resource for genome annotation, microarray development, phylogenetic analysis and other molecular biology applications in entomology.

[Gokhale K, Patil DP, Dhotre DP, Dixit R, Mendki MJ, Patole MS and Shouche YS 2013 Transcriptome analysis of *Anopheles stephensi* embryo using expressed sequence tags. *J. Biosci.* **38** 301–309] **DOI** 10.1007/s12038-013-9320-0

1. Introduction

Morphological patterning during early embryonic stages, including the process of segmentation, has diverged significantly among the insects (Galis *et al.* 2002). One of the key morphological events during embryogenesis is the germ band retraction (GBR) stage. During GBR, the midgut fuses and encloses the yolk sac laterally; the tracheal pit extensions fuse to form the tracheal tree, and the segmental furrows form from the anterior to the posterior end (Schock and Perrimon 2002). The GBR is also accompanied by patterning of major body axes and segmentation (Sander 1976). However, studies regarding the GBR are limited and are carried out in handful of insect lineages (Schock and Perrimon 2002; Blythe *et al.* 2012).

The mosquito belongs to the basal Nematocera branch and has diverged from the *Drosophila* lineage for ~200 million years ago during which they have accumulated variations on the basic dipteran bauplan (Bateman *et al.* 2004; Powers *et al.* 2000). In mosquitoes, the GBR stage occurs 16–22 h after fertilization (Monnerat *et al.* 2002). During this stage, 180º rotation of the embryo along its longitudinal axis brings the dorsal embryo and the ventral egg side together, a process not observed in *Drosophila* (Monnerat

Keywords. *Anopheles stephensi*; cDNA library; germ band retraction; mosquito; transcriptome

Published online: 18 April 2013

et al. 2002). In addition, there are differences in the formation of extra-embryonic membranes between *Drosophila* and 'lower' dipterans. In *Drosophila*, the extra-embryonic membrane is represented by single 'amnioserosa', whereas in most of the insects including mosquito there are distinct 'amnion' and 'serosa' tissues. In order to gain more insight into this process and its possible role in generating morphological diversity, it is necessary to study this process in mosquito at a transcriptome level. To this end, we have constructed, sequenced and characterized 16–22 h post-fertilization cDNA library from *Anopheles stephensi*. The transcriptome consists of several notable transcripts as identified by the GO terms, majorly related to protein synthesis machinery. We also detected an enrichment of diverse transcripts active in the insect metabolism and development. The generated sequences provide a comprehensive resource and framework towards the further studies in mosquitoes and append to the already available (Patil *et al.* 2009; Dixit *et al.* 2009, 2011).

2. Materials and methods

2.1 *Collection of embryos and isolation of RNA from whole embryos*

A. stephensi (NIV strain) population was reared at 28 (±2)°C and 80% (±5%) humidity under 12 h alternating dark/light cycles. 48–72 h after blood feeding, adult female mosquitoes were allowed to lay eggs in tap water under the standard experimental conditions. Freshly laid eggs were collected and maintained at 28°C until the required age. For RNA isolation, fifty *A. stephensi* embryos (16-22 h post-fertilization) were crushed in a RNase-free glass dounce homogenizer in TRIzol (Invitrogen, Carlsbad, CA). Total RNA isolation was further accomplished following the manufacturer's instructions. RNA quantification was done using UV spectrophotometry and integrity was checked using denaturing agarose gel electrophoresis.

2.2 *Construction of cDNA library and sequencing*

A directional cDNA library was prepared using Creator™ SMART™ cDNA library construction kit (BD biosciences, USA) as per manufacturer's instructions using 1 µg of total RNA. *Sfi* I–digested cDNA fragments were size fractionated using CHROMA SPIN-400 columns and the fractions were analysed on 1.5% agarose/EtBr gel. cDNA fragments ranging from 300 bp to 3 kb were pooled together and were ligated to pDNR-LIB vector. Ligation mix was precipitated and reconstituted in distilled water. Electroporation was carried out using GenePulser (Biorad, USA) in DH10B *E. coli* (Invitrogen, Carlsbad, CA) using the supplier's instructions. Transformation mix was spread on pre-warmed LB plates containing chloramphenicol (12.5 µg/mL) and was grown overnight at 37°C to obtain the clones. Over 1500 bacterial clones were randomly picked up and inoculated in LB with chloramphenicol for plasmid preparation. Plasmid extraction was carried out using Montage Plasmid Miniprep$_{96}$ kit (Millipore). The DNA was stored at –20°C until sequencing was carried out. Single-pass 5′ ESTs were generated by sequencing these plasmids using vector-specific M13 forward primer (5′-GTAAAACGACGGCCAGTAGATCT-3′) with BigDye Terminator v3.1 Chemistry (Applied Biosystems, Foster City, CA) on ABI 3730 Genetic Analyzer (Applied Biosystems, Foster City, CA).

2.3 *Sequence analysis*

Base calling was performed using Phred (Ewing *et al.* 1998) (quality≥20) with minimum length greater than 100 bases. Vector, primer and adapter sequences were removed using cross_match. EST clustering was performed using CAP3 (Huang and Madan 1999). Redundancy was estimated by using the formula, $1-((no.\,of\ contigs+no.\,of\ singlets)/total\ \ no.\,sequences.)\times 100$. BLAST analyses were done locally using the tools available at NCBI (*ftp://ftp.ncbi.nih.gov/blast/executables/*) (Altschul *et al.* 1997). Putative functions were assigned to the sequences through comparison against protein (BLASTX, e-value≤ 1e−05) and nucleotide (BLASTN, e-value≤1e−05) databases. GO terms were assigned through gene ontology fasta subset (Lewis *et al.* 2000). Conserved domains were searched using conserved domains database (CDD) of NCBI (Marchler-Bauer *et al.* 2002) containing the KOG (Tatusov *et al.* 2003), Pfam (Bateman *et al.* 2004) and Smart motifs (Schultz *et al.* 2000). BLAST-based comparisons were also made using downloaded databases containing mitochondrial and rRNA nucleotide sequences from NCBI. Additionally, transcript sequences were screened for potential open reading frames (ORF) and the predicted amino acid sequences were analysed for the presence of potential signal peptide cleavage sites on SingalP server (*http://www.cbs.dtu.dk/services/SignalP/*) (Nielsen *et al.* 1997). All ESTs are submitted in NCBI GenBank database with accession numbers FL483337 to FL484476.

3. Results

3.1 *Summary of the EST library from 16–22 h embryos*

During the course of present investigation, we have constructed, sequenced and characterized an embryonic

sequence tag (EST) library from the 16–22 h *Anopheles stephensi* embryos. This stage corresponds to the GBR stage during the embryogenesis and is one of the most morphologically active stages. The genome size of *Anophels stephensi* is ~240 Mbp (Sharakhova *et al.* 2010). With the current estimate that only 1.2% of the genome codes for proteins, our library thus represents ~20% of the transcriptome (total coding potential) of this mosquito species. We obtained 1140 ESTs, which assembled into 1002 unique transcripts with an average size of 412 bp, illustrating a low redundancy (~12.1%) (table 1). The unique sequences are henceforth referred as 'contigs'. More information can be found in table 2 for prevalent EST sequences in some contigs. With the availability of the genomic and EST databases from other well-studied insect species, it has been easier to compare and get quick information about sequences from any unknown genome. Therefore, in order to determine the putative nature and origin, we searched *A. stephensi* embryonic ESTs dataset against other available insect and non-redundant databases, using BLASTX parameters (for other comparisons, refer table 3). As expected, analysis indicated overall maximum 95% homology to mosquito *Anopheles gambiae* while least homology to the mosquito *Culex quinquifaciatus* ESTs. Such information is very important to elucidate the molecular differences, may be either due to average length of the sequences in the library or can actually be novel transcripts previously unidentified due to limited dataset of developmental stages of the mosquitoes.

3.2 *Functional prediction and classification of sequences*

To predict possible functions of the putative transcripts, we first analysed ESTs dataset against multiple protein databases, including non-redundant (NR), GO, Pfam, SMART, KOG, etc. Based on the best hits and e-values (10^{-5}), and putative predicted functional categories, we further classified the ESTs into 6 classes based upon their putative function (table 4; figure 1): I (Energy and metabolism); II (Cytoskeleton and cell structure); III (Intracellular trafficking and vesicular transport); IV (Protein synthesis machinery); V (Signal transduction) and VI (Conserved function). Among the recognized protein coding genes, class IV showed highest fraction of genes (35%), followed by gene categories 'Energy and metabolism' (22%). Genes annotated as 'conserved hypothetical proteins' class represent the third largest fraction of the sequences. It is noteworthy that the most of the sequences in the library are singlet whereby the overall redundancy factor of the library is limited to only 12%. Except for the ESTs classified into class IV (protein synthesis), all of the ESTs are sequenced only once. Thus, the library is a good source of novel/unidentified transcripts in this stage of mosquito development. The remaining major fraction (~70%) of sequences did not yield any significant hits/homology, could be grouped as 'UNKNOWN', demanding further functional analysis through reverse genetic approaches.

Table 1. Summary of 16–22 h embryonic library from *Anopheles stephensi*

Total ESTs analysed	1140
Number of contigs[1]	108
Number of ESTs in contigs[2]	246
Number of singletons[3]	894
Number of putative transcripts[4]	1002
Average contig size (in bases)	412
Redundancy[5]	12.1%

[1] Total sequences assembled from multiple ESTs. [2] EST sequences, which assembled into contigs. [3] These were found in singular copies and were not clustered to in a contig. [4] Number of singletons and contigs together. [5] Transcript redundancy in the library, for calculations refer 'section 2'.

3.3 *Molecular characterization of putative novel transcripts*

(a) Analysis of mosquito *Profilin*: During development process, actin-binding regulatory protein PROFILIN plays an important role in fast cellular architecture remodelling. Our primary analysis of a partial cDNA sequence (297 bp), encoding a 92-amino-acid-long peptide showed significant hits to the previously reported profilin sequences from other insects. Specific pBLAST analysis of the As*Prof* (GenBank: FL483965) revealed the Profilin superfamily domain prediction, while multiple alignment analysis showed high degree of conservation (figure 2a), suggesting the conserved function of profilin in different insect species.

(b) Analysis of *Innexin*-like protein: In the current study, we identified a new putative partial cDNA sequence (215 bp), encoding a 70-amino-acid-long peptide, a homologue of Innexin transmembrane family proteins. The innexin proteins belong to gap junction family members, participating during embryonic development and neural signalling. Multiple alignment analysis showed a high degree of conservation with other insect Innexin proteins, resulting maximum identity of 80% to the mosquito *A. gamabie* and least identity of 38% to the *Bombyx mori*. The conservation pattern of As*Innexin* sequence coverage predicted similarities to the partial cytoplasmic loop (CL) and transmembrane 3 (TM3) domain (figure 2b).

Table 2. Details of ESTs in the assembled contigs

Contig ID	Length	No. of ESTs per contig	GenBank accession No. [2]	SignalP average result[3]	Best match to NR database (NCBI) [4]	E-value[5]	Sequence ID[6]
Contig_67	334	2	FL483993	IND	-	Insignificant	-
Contig_337	142	2	FL483384	CYT	-	Insignificant	-
Contig_184	304	2	FL483917	CYT	-	Insignificant	-
Contig_188	247	2	FL483740	IND	-	Insignificant	-
Contig_577	184	2	FL483527	SIG	-	Insignificant	-
Contig_383	119	2	FL484042	IND	-	Insignificant	-
Contig_239	767	3	FL484467	IND	ENSANGP00000016934	6.00E−60	gb|EAA09966.2
Contig_2	436	3	FL483849	CYT	ENSANGP00000009009	8.00E−45	gb|EAA00360.2
Contig_192	333	3	FL483985	CYT	-	Insignificant	-
Contig_65	1157	3	FL484400	CYT	-	Insignificant	-
Contig_294	560	3	FL484312	CYT	-	Insignificant	-
Contig_173	328	3	FL483968	SIG	-	Insignificant	-
Contig_250	285	3	FL483864	CYT	-	Insignificant	-
Contig_8	442	4	FL484191	CYT	Tar1p	2.00E−09	-
Contig_118	408	4	FL484133	CYT	ribosomal protein L41	2.00E−06	-
Contig_127	342	4	FL484033	CYT	-	Insignificant	-
Contig_14	418	5	FL484150	CYT	ENSANGP00000016990	3.00E−24	gb|EAA09967.3
Contig_242	360	5	FL484054	CYT	-	Insignificant	-
Contig_136	221	5	FL483683	CYT	-	Insignificant	-
Contig_18	626	10	FL484317	CYT	ENSANGP00000028538	7.00E−09	gb|EAL38898.1

[1] Longest EST among the assembled ones into the respective contig. [2] GenBank accession number of the representative EST forming the contig. [3] Potential presence of signalP: SIG, signal-P; CYT, cytoplasmic; Ind, indeterminate. [4] e-value≤1e−05. [5] 'Insignificant' represents a high E-value. [6] GenBank accession number of the blast hit. Note that 'gb' represents GenBank database.

(c) Analysis of putative low temperature viability (LTV) protein: In the present embryonic transcriptomic screen, we identified a cDNA (338 bp), encoding putative LTV-like proteins, known to involved in maintaining genetic co-ordination between stress response and ribosome biogenesis. The translated amino acid sequence showed 94% identity to the unclassified protein from mosquito *A. gamabie*. Multiple amino acid sequence alignment resulted high conservation among the aligned other insect LTV sequences (figure 2c).

Table 3. Summary of organism specific BLAST analysis (e-value≤ 1e−05)

Name of database	Percentage match	BLAST Type
Aedes aegypti	25.45	BLASTX
Anopheles gambiae DNA	60.28	BLASTN
Anopheles gambiae EST	76.05	BLASTX
Anopheles gambiae PEP	23.05	BLASTX
Anopheles gambiae TRANS	30.34	BLASTX
Anopheles stephensi midgut EST	27.85	BLASTX
Culex pipiens quinquefasciatus EST	20.06	BLASTX
Drosophila melanogaster EST	28.24	BLASTX
NR database (NCBI)	21.25	BLASTX

4. Discussion

Early embryonic stages in insects are commonly studied phenotypically by microscopy; however, recent advances in sequencing technologies has allowed to investigate the transcriptome of these stages (Koutsos *et al.* 2007; Biedler *et al.* 2012). We have presented for the first time the transcriptome of germ band stage embryos in *Anopheles stephensi*. We report a total of 1140 ESTs functionally classified into 6 classes. Majority of the transcripts do not show match with the known transcripts in the database. Some of these transcripts could be mosquito specific and would be annotated as a result of further refinement of the genomic sequencing efforts.

Class I constitutes all the sequences functionally annotated in the 'Energy and metabolism' class. The most abundant sequences belong to enzymes that are components of various metabolic pathways. In oviparous animals, embryogenesis

Table 4. Distribution of ESTs analysed from 16-22 h *A. stephensi* embryos

Class	Gene function category	Number of EST sequences	Number of unique transcripts or contigs	Redundancy factor	Expression percentage
Class I	Energy and metabolism	48	48	1	22
Class II	Cytoskeleton and cell structure	27	27	1	13
Class III	Intracellular trafficking	8	8	1	4
Class IV	Protein synthesis	129	72	1.79	34
Class V	Signal transduction	20	20	1	9
Class VI	General function	38	38	1	18

ESTs showing match with the NR database (e-value≤1e−05) were considered for analysis.

occurs in absence of any exogenous nutritional supply. In such a scenario, nutrients are accumulated in the oocytes during the process of oogenesis. It has been shown recently that the germ band formation stage is a landmark regarding both glucose and glycogen metabolism (Vital *et al.* 2010), wherein glucose metabolism was investigated throughout the embryonic development of *A. aegypti*. The authors report an increase in the glycolytic pathways after the germ band formation in *A. aegypti* embryos (Vital *et al.* 2010). In the present library, transcripts FL484176 (6-phosphofrucokinase), FL484021 (phosphoglycerate kinase) and FL484394 (pyruvate dehydrogenase) are all members of the glycolytic pathway, corroborating the finding that glycolysis is intensified at the end of GBR stage.

Embryogenesis is a complex process and involves large-scale cell migration and reorganization. Among the transcripts coding for cytoskeleton and cell structure components (class II), we highlight the presence of transcript coding for a membrane-associated guanylate kinase (MAGUK) family protein (FL484404), which is implicated to play an important role in the establishment and maintenance of polarity of embryonic epithelium (Caruana 2002).

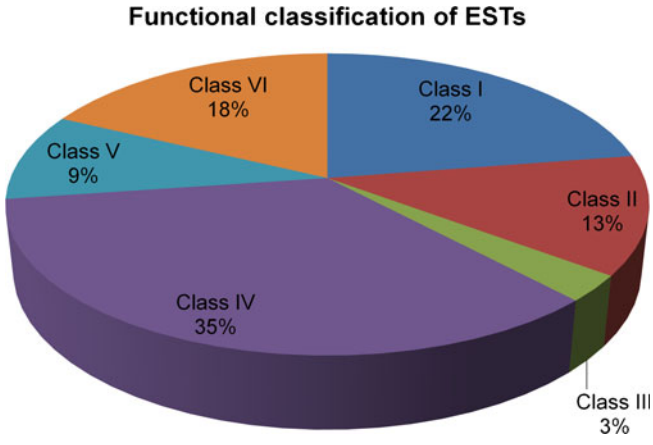

Figure 1. Functional classification of ESTs. Sequences were classified by GO terms. Sequences with a significant match with the NR database (e-value≤1e−05) were considered for this analysis.

Proteins belonging to the MAGUK family are multidomain scaffolding proteins and are present on the plasma membrane of the polarized epithelial cells. Genetic studies in *Drosophila* and *C. elegans* have implicated the role of these proteins in cell junction assembly and organization of polarized signaling complex. Recent studies have shown this protein to be associated with several biological processes including regulation of cell polarity, asymmetric cell division and cell invasion (Robert *et al.* 2012). The presence of this transcript from the GBR stage embryos in mosquitoes suggests conserved mechanisms for regulation of cell polarity. As described above, during the GBR stage, a 180º rotation of the embryo along the longitudinal axis brings together the dorsal and ventral side. This process is not observed in any other dipteran and is thought to be a unique feature of the mosquito embyrogenesis (Valle *et al.* 1999). Profilin (actin-regulating protein-chikadee) has been implicated in gastrulation movements and play component roles in directing filopodial protrusion and other aspects of epithelial and amnioserosa cell behaviour in *Drosophila* (Jacinto *et al.* 2002). Transcript FL483965 from our data is mosquito homologue of the profilin gene. It would be interesting to further examine and compare the role of these genes in the establishment and maintenance of epithelial cell polarity in this mosquito and other insects. The actin cytoskeleton is an integral component of all cells and actin modelling is vital to all cellular motility, adhesion and contraction. GBR stage embryos undergo extensive morphological activity and actin regulation must be an important feature of this process. We have identified some transcripts encoding actin regulators in the library. *Flightless I* gene is highly conserved across invertebrates and vertebrates and is speculated to be a part of a signal transduction pathway that links cytoskeleton structure with transcriptional regulation (Li *et al.* 2004). We have identified putative homologues of Flightless I (FL484405) as well as Rab6 (FL484226), both are implicated in the regulation of actin cytoskeleton (Coutelis and Ephrussi 2007).

Only 3% of the transcripts were classified as those playing role in intracellular trafficking and vesicular transport (class III). Notable among these transcripts are those

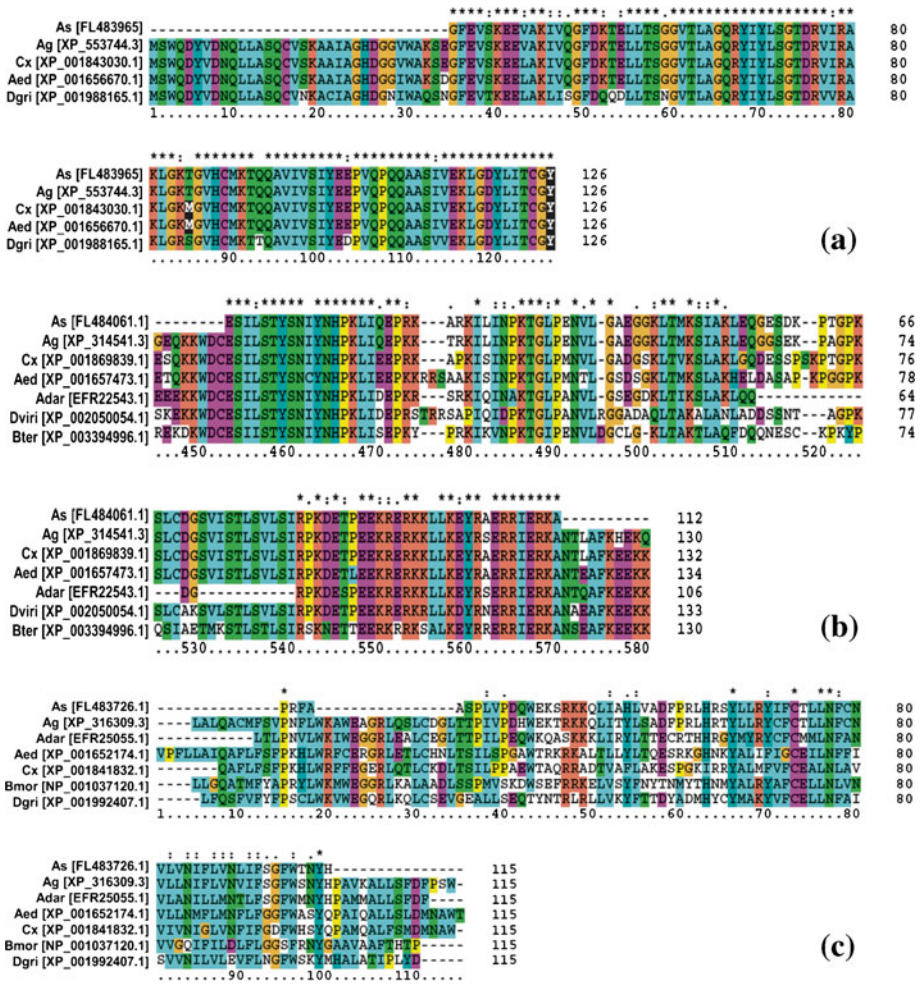

Figure 2. Molecular analysis of new putative cDNA transcripts, identified from mosquito *A. stephensi*, GBR stage transcriptome. (**a**) Multiple sequence alignment of *AsProfilin* with other insect profilins. (**b**) Multiple sequence alignment of *AsInnexin* with other insect Innexins. (**c**) Multiple sequence alignment of *AsLTV* with other insect LTVs. Conserved residues are asterisked. Gene bank accession numbers are indicated label of each sequence.

coding for GAP-junction-forming Innexin family protein (FL484321) and a NIEMANN pick type C2 protein (NPC2-related) (FL483726), both showing homology to ENSANGP00000020577 and ENSANGP00000017003, respectively. The NPC2 gene encodes a protein that binds and transports cholesterol. Cholesterol is deposited in the eggs to meet the regulatory and metabolic needs of the embryo during its development. As described above, in the germ band formation stage, there is a rise in the process of gluconeogenesis, which is result of breakdown of noncarbohydrate precursors such as lipids (Vital *et al.* 2010). It can be speculated that this protein is implicated in the binding and transfer of cholesterol during this stage of development. Although a *Drosophila* homologue of the protein is implicated in sterol biosynthesis and homeostasis, the expression profile of the gene remains uncharacterized. The detection of this transcript indicates similar metabolic fate in all the insects during the GBR stage.

Highest number of transcripts showed matches with genes that take part in transcription, translation, RNA processing and chromatin structure and dynamics. The GBR stage is accompanied by a transition from parasegmental to segmental division of the embryo and the embryo is transcriptionally active in this stage of development. Transition of parasegmental to segmental division of embryo is controlled by crosstalk between the homeobox genes, which in turn undergo a complex regulation by the members of the polycomb (PcG) and trithorax group (TrxG) of proteins. We are able identify members from the TrxG proteins. Contig 644 (FL483522) shows homology to the *Drosophila toutatis* gene. The *toutatis* gene is ubiquitously expressed during all stages in development and

plays important roles in regulation of transcription by RNA polymerase and in wing development through chromatin remodelling (Emelyanov *et al.* 2012). We also identified a homologue of Histone H3 (Lys9) methyltransferase SUV39H1/Clr4 (FL484397). Homologue of the protein adds methyl group to lysine 9 of histone H3, then recognized by heterochromatic protein HP1, which causes gene silencing (Bannister *et al.* 2001). Expression of these genes hint at conserved pathways of epigenetic regulation associated with GBR stage embryos. Developmental timing of gene expression is regulated by a set of evolutionarily conserved heterochronic genes, which specify the cell fate in a stage specific manner. An example of such genes is the RNA-binding gene *lin28*. Lin 28 protein contains two RNA binding domains: the cold shock domian (CSD) and a pair of retroviral-type CCHC zinc fingers (Moss and Tang 2003). In *Drosophila*, *lin28* expression is observed in twp stages, first during embryogenesis through the first instar larvae and in the pupal stage (Moss and Tang 2003). We also have been able to identify the lin28 homologue (FL484372) in our library. One of the most abundant transcripts in this class is homologue of Pur-alpha (FL484150). Pur-alpha is sequence specific single stranded nucleic acid binding protein and is highly conserved across animal phyla. Pur-alpha is implicated in diverse cellular functions, including transcriptional activation and repression, translation and cell growth (Gallia *et al.* 2000). Although Pur-alpha is shown to have several interacting partners and have diverse roles, in the present stage of development, it would most likely be interacting with Cyclin A and regulating cell growth. Mosquitoes show good adaptability to the changing environment and this ability is one of the reasons implicated in their widespread distribution. Such adaptation processes are inherently coupled to changes in the expression of functional proteins that, in turn, are based on regulation of molecular processes such as transcription initiation, mRNA stability, translation or posttranslational protein modification. The Ccr4-Not complex acts on most of these processes to control basic proteins expression and thereby plays a key role in adaptive responses to environmental challenges (Lenssen *et al.* 2005). The complex contains CCR4, Not1–5 domains as well as some other proteins such as CAF1, CAF40 and CAF 130. We were able to identify two homologous components of this complex, Not1 (FL484065) and Not5 (FL484159). Not1 is the largest and the only essential component of this complex and it is unlikely that it would have any divergent function. Not5, on the other hand, is shown to contact multiple components of the TFIID complex, which itself facilitates promoter responses to various activators and repressors. Not5 has been shown to coordinate and act in parallel to cell stress pathways by associating with different promoters in a stress-dependent fashion. In this context, we also wish to highlight the presence of transcript FL484061, a homologue of a Low LTV protein isolated from *S. cerevisieae*, which has homologues across the animal kingdom. The exact function of the protein remain unknown; however, yeast strains lacking this gene reveal an unknown link between ribosome biogenesis factors and environmental stress sensitivity (Loar *et al.* 2004). It would be interesting to study if there is any interaction between Not5 and LTV that in turn would have any implications on the stress response mechanisms in mosquito embryogenesis and life cycle in general. In addition to above transcripts, we observe that the library represents homologues of transcriptional repressors like transcriptional corepressor Atrophin-1/DRPLA (FL484095) and CBF1-interacting corepressor CIR (FL484126). Prevalence of several repressors indicates their active and complex role in the regulation of gene expression during this stage of the insect life cycle.

In embryonic development, signalling networks act to achieve cell fate specification. In the current library, 9% of the sequences belong to components/elements from the signal transduction pathways. The presence of ubiquitin domain sequences (FL483388 and FL483417) further corroborates the role of ubiquitination in mosquito development as reported previously (Koutsos *et al.* 2007). In addition to these, the library also represents a homologue (FL483715) of odorant-binding protein OBP-9 in mosquito *A. gambiae* and its counterpart in *Drosophila* (Obp44a). OBP-9 belongs to class of atypical odorant-binding proteins, counterparts of which have been implicated in non-olfactory functions. However, based on the sequence similarity with other genes in *Drosophila*, this transcript has been grouped in the canonical Obp gene cluster which shows olfactory system expression (Hekmat-Scafe *et al.* 2002). The presence of this transcript in embryonic stage is very intriguing. It is possible that the protein is subverted to another function like members of the atypical odorant-binding proteins. Other notable transcripts include homologs for defensin (FL484397). Defensins are small cysteine rich proteins found in vertebrates as well as invertebrates and plants. Defesins contain a signature knottin domain, which also represents plant lecithins, recently been reported from *A. stephensi* salivary gland (Dixit *et al.* 2008). Another notable transcript encodes for an adapter protein Disabled (FL484215). Disabled is required during signalling by the sevenless receptor protein tyrosine kinase and also functions towards the identification of additional DRK-binding proteins (Le and Simon 1998).

Third largest fraction of the library (Class VI-18%) constitutes transcripts classified into the ‘General function prediction/function unknown’ category. Several of them are involved in conserved processes such as RNAi and may be other non-coding RNAs whose functions are yet to be discovered. Others are associated with several functions such as plasma membrane biogenesis (FL483649) and structural constituents of ribosomes (FL 484041, FL483853, FL483606 and FL484583). Some of the transcripts are also associated with transcription related processes. Transcripts

FL484000 and FL484049 are associated with RNA binding, while FL484065 is classified into the 'General function prediction' category and is implicated as negative regulator of transcription. Some of the transcripts (FL484192, FL484178, FL484280, FL483963 and FL484117) are implicated in various metabolic processes.

5. Conclusions

GBR is one of the important stages during insect development. In addition to being implicated with considerable epithelial morphogenesis, the stage may also play an important role in the body plan specification in insects. However, studies regarding the GBR transcriptome are limited. We have constructed and partially sequenced a cDNA library from *Anopheles stephensi* GBR stage embryos. The generated sequences in this work provide a comprehensive resource for genome annotation, microarray development, phylogenetic analysis and other molecular biology applications in mosquito embryonic stages. Comparative analysis across sequences from this stage in other insects might also shed light on the potential mechanism responsible for generating diversity in animal forms.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25** 3389–3402

Bateman A, Coin L, Durbin R, Finn R D, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, Studholme DJ, Yeats C and Eddy S R 2004 The Pfam protein families database. *Nucleic Acids Res.* **32** D138–D141

Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC and Kouzarides T 2001. Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410** 120–124

Biedler JK, Hu W, Tae H and Tu Z 2012 Identification of early zygotic genes in the yellow fever mosquito *Aedes aegypti* and discovery of a motif involved in early zygotic genome activation. *PLoS One* **7** e33933

Blythe MJ, Malla S, Everall R, Shih YH, Lemay V, Moreton J, Wilson R and Aboobaker AA 2012 High through-put sequencing of the Parhyale hawaiensis mRNAs and microRNAs to aid comparative developmental studies. *PLoS One* **7** e33784

Caruana G 2002 Genetic studies define MAGUK proteins as regulators of epithelial cell polarity. *Int. J. Dev. Biol.* **46** 511–518

Coutelis JB and Ephrussi A 2007 Rab6 mediates membrane organization and determinant localization during Drosophila oogenesis. *Development* **134** 1419–1430

Dixit R, Sharma A, Patole MS and Shouche YS 2009 Salivary gland transcriptome analysis during Plasmodium infection in malaria vector *Anopheles stephensi*. *Int. J. Infect. Dis.* **13** 636–646

Dixit R, Sharma A, Patole MS and Shouche YS 2008 Molecular and phylogenetic analysis of a novel salivary defensin cDNA from malaria vector *A. stephensi*. *Acta Tropica.* **106** 75–79

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

Emelyanov AV, Vershilova E, Ignatyeva MA, Pokrovsky DK, Lu X, Konev AY and Fyodorov DV 2012 Identification and characterization of ToRC, a novel ISWI-containing ATP-dependent chromatin assembly complex. *Genes Dev.* **26** 603–614

Ewing B, Hillier L, Wendl MC and Green P 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8** 175–185

Galis F, van Dooren TJ and Metz JA 2002 Conservation of the segmented germband stage: robustness or pleiotropy? *Trends Genet.* **18** 504–509

Gallia GL, Johnson EM and Khalili K 2000 Puralpha: a multifunctional single-stranded DNA- and RNA-binding protein. *Nucleic Acids Res.* **28** 3197–3205

Hekmat-Scafe DS, Scafe CR, McKinney AJ and Tanouye MA 2002 Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. *Genome Res.* **12** 1357–1369

Huang X and Madan A 1999 CAP3: A DNA sequence assembly program. *Genome Res.* **9** 868–877

Jacinto A, Woolner S and Martin P 2002 Dynamic analysis of dorsal closure in Drosophila: from genetics to cell biology. *Dev. Cell* **3** 9–19

Koutsos AC, Blass C, Meister S, Schmidt S, Maccallum RM, Soares MB, Collins FH, Benes V, Zdobnov E, Kafatos FC and Christophides GK 2007 Life cycle transcriptome of the malaria mosquito Anopheles gambiae and comparison with the fruitfly *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **104** 11304–11309

Le N and Simon MA 1998 Disabled is a putative adaptor protein that functions during signaling by the sevenless receptor tyrosine kinase. *Mol. Cell Biol.* **18** 4844–4854

Lenssen E, James N, Pedruzzi I, Dubouloz F, Cameroni E, Bisig R, Maillet L, Werner M, Roosen J, Petrovic K, Winderickx J, Collart MA and De VC 2005 The Ccr4-Not complex independently controls both Msn2-dependent transcriptional activation–via a newly identified Glc7/Bud14 type I protein phosphatase module–and TFIID promoter distribution. *Mol. Cell Biol.* **25** 488–498

Lewis S, Ashburner M and Reese M G 2000 Annotating eukaryote genomes; *Curr. Opin. Struct. Biol.* **10** 349–354

Li YH, Campbell HD and Stallcup MR 2004 Developmentally essential protein flightless I is a nuclear receptor coactivator with actin binding activity. *Mol. Cell Biol.* **24** 2103–2117

Loar JW, Seiser RM, Sundberg AE, Sagerson HJ, Ilias N, Zobel-Thropp P, Craig EA and Lycan DE 2004 Genetic and biochemical interactions among Yar1, Ltv1 and Rps3 define novel links between environmental stress and ribosome biogenesis in *Saccharomyces cerevisiae*. *Genetics* **168** 1877–1889

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY and Bryant SH 2002 CDD: a database of conserved

domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.* **30** 281–283

Monnerat AT, Machado MP, Vale BS, Soares MJ, Lima JB, Lenzi HL and Valle D 2002 Anopheles albitarsis embryogenesis: morphological identification of major events. *Mem. Inst. Oswaldo Cruz* **97** 589–596

Moss EG and Tang L 2003 Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. *Dev Biol.* **258** 432–442

Nielsen H, Engelbrecht J, Brunak S and von Heijne G 1997 Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engr.* **10** 1–6

Patil DP, Atanur S, Dhotre DP, Anantharam D, Mahajan VS, Walujkar SA, Chandode RK, Kulkarni GJ, *et al*. 2009 Generation, annotation, and analysis of ESTs from midgut tissue of adult female Anopheles stephensi mosquitoes. *BMC Genomics* **10** 386

Powers TP, Hogan J, Ke Z, Dymbrowski K, Wang X, Collins FH and Kaufman TC 2000 Characterization of the Hox cluster from the mosquito *Anopheles gambiae* (Diptera: Culicidae). *Evol. Dev.* **2** 311–325

Robert S, Delury C and Marsh E 2012 The PDZ protein discs-large (DLG): the 'Jekyll and Hyde' of the epithelial polarity proteins. *FEBS J.* **279** 3549–3558

Sander K 1976 Specification of the basic body pattern in insect embryogenesis. *Adv. Insect Physiol.* **12** 125–238

Schock F and Perrimon N 2002 Cellular processes associated with germ band retraction in Drosophila. *Dev. Biol.* **248** 29–39

Schultz J, Copley RR, Doerks T, Ponting CP and Bork P 2000 SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28** 231–234

Sharakhova MV, Xia A, Tu Z, Shouche YS, Unger MF and Sharakhov IV 2010 A physical map for an Asian malaria mosquito, *Anopheles stephensi*. *Am. J. Trop. Med. Hyg*. **83** 1023–1027

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, *et al.* 2003 The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* **4** 41

Valle D, Monnerat AT, Soares MJ, Rosa-Freitas MG, Pelajo-Machado M, Vale BS, Lenzi HL, Galler R and Lima JB 1999 Mosquito embryos and eggs: polarity and terminology of chorionic layers. *J. Insect Physiol.* **45** 701–708

Vital W, Rezende GL, Abreu L, Moraes J, Lemos FJ, Vaz Ida S Jr and Logullo C 2010 Germ band retraction as a landmark in glucose metabolism during Aedes aegypti embryogenesis. *BMC Dev. Biol.* **10** 25

MS received 02 August 2012; accepted 07 March 2013

Corresponding editor: SUDHA BHATTACHARYA