OMICS A Journal of Integrative Biology
Volume 18, Number 7, 2014

DOI: 10.1089/omi.2014.0007

Original Articles

Brain Proteomics of *Anopheles gambiae*

Sutopa B. Dwivedi,[1,2] Babylakshmi Muthusamy,[1,3] Praveen Kumar,[1] Min-Sik Kim,[4] Raja Sekhar Nirujogi,[1,3] Derese Getnet,[4] Priscilla Ahiakonu,[5] Gourav De,[1,6] Bipin Nair,[2] Harsha Gowda,[1] T.S. Keshava Prasad,[1–3,6] Nirbhay Kumar,[7] Akhilesh Pandey,[1,4] and Mobolaji Okulate[5]

Abstract

Anopheles gambiae has a well-adapted system for host localization, feeding, and mating behavior, which are all governed by neuronal processes in the brain. However, there are no published reports characterizing the brain proteome to elucidate neuronal signaling mechanisms in the vector. To this end, a large-scale mapping of the brain proteome of *An. gambiae* was carried out using high resolution tandem mass spectrometry, revealing a repertoire of >1800 proteins, of which 15% could not be assigned any function. A large proportion of the identified proteins were predicted to be involved in diverse biological processes including metabolism, transport, protein synthesis, and olfaction. This study also led to the identification of 10 GPCR classes of proteins, which could govern sensory pathways in mosquitoes. Proteins involved in metabolic and neural processes, chromatin modeling, and synaptic vesicle transport associated with neuronal transmission were predominantly expressed in the brain. Proteogenomic analysis expanded our findings with the identification of 15 novel genes and 71 cases of gene refinements, a subset of which were validated by RT-PCR and sequencing. Overall, our study offers valuable insights into the brain physiology of the vector that could possibly open avenues for intervention strategies for malaria in the future.

Introduction

ANOPHELES GAMBIAE MOSQUITO is one of the major vectors responsible for transmission of malaria. Because of the emerging resistance of mosquitoes to insecticides, it is necessary to continue the search for alternatives to complement current malaria intervention strategies. Such efforts require detailed molecular studies to understand the insect biology and their behavior responses. There have been some research studies particularly targeting chemical messaging systems such as the olfaction and other sensory systems that influence the locomotory, host localization, feeding, and mating behavior of the vector. These studies have been largely restricted to antennae and the reproductive organs of *An. gambiae*. In this respect, we adopted a proteomic approach to gain further insights into the chemical communication and other neuronal processes occurring in the brain of the malaria vector.

Mosquitoes have well-developed neuroendocrine systems and their brains contain hormone-producing neurosecretory cells that regulate growth and development. Therefore, a brain study of the vector will identify several classes of proteins associated with neuronal functioning, synapses, and sensory processes, which may affect the survival and transmission ability of the mosquito. *Anopheles gambiae* is known for its ability to respond to chemical messengers for host location, selection, and feeding behavior (Carey et al., 2010). There have been studies on characterizing the odorant receptor proteins and chemosensory genes to elucidate the olfactory signaling mechanisms in mosquitoes (Liu et al., 2010; Xia et al., 2008). Using bioinformatic sequence homology-based approaches, 276 G-protein coupled receptors have been reported, which include 79 olfactory receptors, 76 gustatory receptors, and 66 genes encoding odorant binding proteins in the *An. gambiae* genome. These receptors play a central role in almost all the sensory and chemical

[1]Institute of Bioinformatics, International Technology Park, Bangalore, Karnataka, India.
[2]School of Biotechnology, Amrita Vishwa Vidyapeetham University, Amritapuri, India.
[3]Centre of Excellence in Bioinformatics, School of Life Sciences, Pondicherry University, Puducherry, India.
[4]McKusick-Nathans Institute of Genetic Medicine, Department of Biological Chemistry, Oncology and Pathology, Johns Hopkins University School of Medicine, Baltimore, Maryland.
[5]Department of Natural Sciences, University of Maryland Eastern Shore, Princess Anne, Maryland.
[6]Manipal University, Madhav Nagar, Manipal, Karnataka, India.
[7]Department of Tropical Medicine, Tulane University School of Public Health and Tropical Medicine, New Orleans, Louisiana.

communication-related pathways in the life cycle of a mosquito (Hill et al., 2002). A recent proteomic study focused on the identification of soluble odorant binding proteins (OBP) and chemosensory proteins (CSP) in the male and female antennae of *An. gambiae* (Mastrobuoni et al., 2013). The data reported 24 OBPs, and as expected, female antennae showed increased numbers and abundance as compared to male antennae, while the reverse situation was observed in case of CSPs. Odorant binding and chemosensory proteins direct many important behaviors that are responsible for the vectorial ability of the mosquitoes such as host preference and localization, flight activity, and others. Based on this study, we hypothesize that the brain proteome study will not only reveal olfactory proteins but also identify repertoire of other GPCRs for which no experimental evidence has been reported as of yet in *An. gambiae*. Similarly, neuropeptidomic studies have been conducted on *Aedes aegypti* in order to identify putative bioactive neuropeptides from the brain and other tissues (Predel et al., 2010).

Other brain proteome studies have been conducted on the fleshfly *Sarcophaga crassipalpis* (Pavlides et al., 2011), aphid (Hummon et al., 2006; Huybrechts et al., 2010), the silkworm *Bombyx mori* (Li et al., 2007; Li et al., 2010; Li et al., 2009), and the honey bee (Uno et al., 2007; Wolschin et al., 2009). A differential proteomic study of diapausing and nondiapausing pupal brains of fleshfly have shown higher amounts of stress-related proteins but lower abundance of other metabolism-related brain proteins in the diapausal state (Li et al., 2007). A recent study on the honeybee brain proteome revealed a total of 2742 proteins identified from worker and nurse populations (Hernandez et al., 2012), which were reported to be involved in various biological processes, with 10% that could not be assigned any functions. The data reported the relative abundance of proteins between nurse and worker honey bees and showed an enrichment of specific pathways respective to the two classes of honeybees.

Accurate cataloging of all protein-coding genes in any genome is a vital step towards the discovery of their molecular functions (Lin et al., 2007). Many groups have contributed towards characterizing the proteome of *An. gambiae*. Some of the earliest proteomics efforts included analysis of salivary glands (Kalume et al., 2005) and midgut peritrophic matrix (Dinglasan et al., 2009) that led to identification of 69 and 209 proteins, respectively. Subsequent proteomic efforts have included protein identifications from cast cuticles (He et al., 2007) and eggshell (Amenya et al., 2010) of *An. gambiae*. Analysis of the mosquito hemolymph by two-dimensional gel electrophoresis led to identification of 28 proteins (Paskewitz et al., 2005). Differential proteomics studies have determined the antenna profiles of the male and female mosquitoes through MALDI-based mass spectrometry studies (Dani et al., 2008).

Mass spectrometry-based proteomic methodologies has emerged as a powerful tool, not only to perform large-scale analysis of proteins from complex tissues, characterize protein–protein interactions, but also as a complementary approach to genome annotation of any existing or newly sequenced genomes, which is referred to as proteogenomics. In simpler terms, ''proteogenomics'' can be defined as using peptide data derived from mass spectrometry-based proteomics to refine the annotation of protein coding genes in the genome. The established methods of carrying out genome annotation of a genome sequence include de novo gene prediction, comparative genomics, and experimental evidence-based genome annotation. Each of these methods have their own biases such as computational model based predictions which rely on some signals in the genome or limitations as in comparative genomics where only orthologous genes can be identified based on evolutionary conservation.

Proteogenomics strategy stands out as an important experimental tool to identify the protein coding potential of sequenced or unsequenced genomes of an organism (Castellana et al., 2010; Krug et al., 2011). Proteogenomically identified peptide data can provide invaluable information for gene annotation, which is almost impossible or difficult to predict using nucleotide sequence information alone. Examples include validation or confirmation of an annotated protein coding region, refinement of existing gene models, confirmation or identification of alternate splice forms, and most importantly, identification of novel genes in an uncharacterized genomic location (Hernandez et al., 2014). Proteogenomic analysis in which peptide data is used to revise the genome annotation of organisms is getting more popular (Kalume et al., 2005a, 2005b; Okulate et al., 2007; Pandey et al., 2000). Re-annotation of the *An. gambiae* genome using mass spectrometric-based approaches has been reported (Chaerkady et al., 2011; Holt et al., 2002).

Thus far, no proteomic investigation has been reported for the brain tissue of *An. gambiae*. In this study, we have performed mass spectrometry-based brain proteome profiling, highlighted the roles of several brain proteins involved in important biological processes and signaling pathways, and utilized the peptide data for genome annotation. Our proteogenomic analysis led to identification of 15 novel protein coding genes, 25 cases of N/C-terminal gene extensions, correction of reading frames of 4 gene models, 10 alternate splice forms, and deduced the translational start sites for 10 protein coding genes. In addition, 25 other types of gene modifications have been reported involving novel exons, exon extensions, and UTR translations. A subset of the data has been also validated by c-DNA sequencing. During the course of our proteogenomic analysis, AgamP3.7 version was released by Ensembl, which added 140 new protein coding genes. The findings of our study will provide a framework for further molecular studies on *An. gambiae* brain proteins, particularly those involving genes that control behavior, growth, and development. Such information may prove useful as we continue to search for targets that complement those already in existence for malaria vector control.

Materials and Methods

Sample collection and processing

An. gambiae mosquitoes (G-3 strain collected from MR4) were reared in insectary at 27°C ± 1°C and 80% ± 5% relative humidity and 12 h light/dark cycle. Adults were fed on 10% pancake syrup. Brain tissues from 500 adult mosquitoes (male/female) were dissected using an Olympus SZX12 stereomicroscope and stored at −80°C until LC-MS/MS analysis.

Sample fractionation

The brain tissues from 500 mosquitoes were lysed in 4% SDS with 100 mM Tris HCl buffer at pH 8.5. The lysed

tissues were boiled for 2–4 min at 95°C. Protein concentration was estimated using Bio-Rad DC protein assay. Approximately 260 μg of protein from pooled brain sample was resolved using 10% SDS-PAGE (16 × 18 cm gel). The gel was stained with colloidal Coomassie blue (Invitrogen, Carlsbad, CA). Thirty-three gel bands were excised, destained, and subjected to in-gel trypsin digestion as previously described (Harsha et al., 2008). Briefly, the excised bands were reduced with 5 mM dithiothreitol in 40 mM ammonium bicarbonate, followed by alkylation with 10 mM iodoacetamide in 40 mM ammonium bicarbonate. Digestion was carried out using trypsin (modified sequencing grade; Promega, Madison, WI) at 37°C for 16 h. The peptides were extracted from the gel slices as described in earlier studies (Harsha et al., 2008), dried and stored at −80°C until LC-MS/MS analysis was carried out.

LC MS/MS analysis

Nanoflow electrospray ionization tandem mass spectrometric analysis of peptide samples was carried out using LTQ-Orbitrap Velos mass spectrometer (Thermo Scientific, Bremen, Germany) interfaced with Easy-nLC II nano flow liquid chromatography system (Thermo Scientific, Odense, Southern Denmark). The chromatographic capillary columns used were packed in-house with Magic C_{18} AQ (Michrom Bioresources, Inc., Auburn, CA; 5 μm particle size, pore size 100 Å) reversed phase material in 100% acetonitrile at a pressure of 1000 psi. The peptide sample from each in-gel digested fraction was enriched on a precolumn (75 μm × 2 cm) at a flow rate of 5 μL/min with solvent A (0.1% formic acid in water). Peptides were separated on an analytical column (75 μm × 10 cm) at a flow rate of 350 nL/min using a linear gradient of 7%–30% solvent B (0.1% formic acid in 95% acetonitrile) over 60 min. Mass spectrometry analysis was carried out in a data-dependent manner with full scans within 350–1800 m/z acquired using an Orbitrap mass analyzer at a mass resolution of 60,000 at 400 m/z. For each duty cycle, twenty most intense precursor ions from a survey scan were selected for MS/MS and detected at a mass resolution of 15,000 at m/z of 400, also in an Orbitrap analyzer. The fragmentation was carried out using higher-energy collision dissociation (HCD) with 39% normalized collision energy. Dynamic exclusion was set to 30 seconds with a 10 ppm mass window. The automatic gain control for full FT MS was set to 0.5 million ions and for FT MS/MS was set to 0.1 million ions with a maximum ion injection times of 100 ms and 200 ms, respectively. Internal calibration was achieved using lock-mass from ambient air (m/z 445.1200025) as described previously (Olsen et al., 2005). Other parameters include spray voltage of 2.0 kV, and capillary voltage of 250.

Bioinformatics analysis

Construction of hypothetical protein database. The genome of *An. gambiae* (AgamP3 assembly released in February 2006) was downloaded from VectorBase. The genome was translated in all the six reading frames using the standard genetic code. The translated sequences from stop codon to the next stop codon were fetched, and a six frame translated genome database was constructed. Such sequences that are less than six amino acids were discarded. This database contained 15,885,149 sequences. Most of the commonly occurring contaminants such as trypsin, keratins, and albumin sequences were added to the translated genome database that was used for MS/MS ion search. The MS/MS data was searched against this six frame translated genome database using Sequest and Mascot search engines to identify potential novel events.

Construction of translated EST sequence database. A total of 236,004 *An. gambiae* ESTs were downloaded from dbEST, Core Nucleotide database. All the ESTs along with the mRNA sequences of AgamP3.6 were translated in three reading frames and used for searching the MS/MS data. The common contaminants were appended to the database.

Construction of hypothetical N-terminal protein database. An N-terminal protein database was generated in house by creating a peptide library of all methionine-containing peptides until the next K/R from the annotated Ensembl and SNAP prediction models. Also, the 5′UTRwas translated in the reading frame of the first coding exon of the Ensembl models and any methionines in the translated sequence was fetched until the next K/R. One missed cleavage was allowed. The database was used for identifying protein N-terminal acetylated peptides using Mascot search engine. The common contaminants were appended to the database.

Construction of *ab initio* predicted protein database. SNAP *ab initio* gene prediction dataset of *An. gambiae* was downloaded from Ensembl using Biomart version 0.7 which was used to construct a predicted protein database of 24,679 sequences. The common contaminants were appended to the database. The MS/MS data was searched against this predicted protein database using Sequest search engine to identify potential novel events.

Data search methods

The mass spectrometry data was processed using Proteome Discoverer (version 1.3) software (Thermo Scientific) workflow with Sequest and Mascot search algorithms against *An. gambiae* (AgamP3.6 genebuild) protein database containing 12,670 sequences and a hypothetical protein database derived from the six frame translation of *An. gambiae* genome. The processed MS/MS spectra were also searched against SNAP *de novo* gene prediction database (a set of predicted gene models by gene prediction algorithm) and database of hypothetical N-terminal peptide sequences. A searchable three frame translated transcript sequence database was created in house with EST and mRNA sequences downloaded from NCBI Core nucleotide database. The search parameters used were oxidation of methionine and protein N-terminal acetylation as variable modifications (protein N-terminal acetylation modification was used only in Mascot protein searches) and carbamidomethylation of cysteine residues as a fixed modification. A maximum of one missed cleavage was allowed for tryptic peptides. The peptide and protein data were extracted using high peptide confidence and top one peptide rank filters. Mass error window of 20 ppm and 0.1 Da were allowed for MS and MS/MS, respectively. 1% FDR was used as a cut-off value for reporting identified peptides.

Workflow for proteogenomic analysis

MS/MS data was searched against 6 databases—Protein database (using Sequest and Mascot), SNAP prediction protein database (using Sequest), six frame translated genome database (Sequest, Mascot), three frame translated EST database (using Sequest), and hypothetical N-terminal database (using Mascot). Peptides obtained after applying 1% FDR cut off were selected for genome annotation analysis. This was done to ensure that our analysis is based only on high confident peptide identification. The peptides that mapped to annotated Ensembl protein coding genes were filtered out for the proteogenomic analysis. The remaining peptides that did not map to any annotated ORFs (GSSPs) were classified based on the following criteria: 1. peptides which mapped to a unique or single region in the genome; 2. The peptides which either mapped to regions where no Ensembl genes were annotated or where the annotations were not in concordance with the peptide data.

Out of the 183 novel peptides hitting unique places in the genome, 37 of them now have protein evidence in the Ensembl AgamP3.7 released in October 2012. Only GSSPs that exhibited a 100% blast alignment to a single region in the *An. gambiae* genome were used for the proteogenomics study. The GSSPs were mapped to the *An. gambiae* genome using Ensembl Blast tool. Based on their genomic location they were categorized into: intergenic peptides, intronic peptides, exon–intron junctional peptides, CDS–UTR junctional peptides, exon–no gene junctional peptides, and exon–exon junctional peptides. Proteogenomic analysis of these GSSPs will therefore aid in identifying novel open reading frames, refine existing gene annotations, and identify alternative splicing events. A complete overview of the workflow adopted for the brain MS/MS data analysis is shown in Figure 1.

RNA isolation and primer designing for RT-PCR

Gene model specific primers were designed using Primer3 (v. 0.4.0) software (Rozen et al., 2000). The primer design was based on annotation of gene models as predicted by gene prediction algorithms or homology to other related species. Primers spanning predicted introns were chosen wherever possible except for single exon genes, mainly to distinguish between genomic DNA and cDNA template. The amplicon size was chosen to be greater than 150 bp to ensure good sequencing quality. Total RNA was isolated from the brain of *An. gambiae* using Qiagen RNeasy kit. The RNA yield was estimated using Nanodrop. cDNA prepared by reverse transcription of RNA using Clonetech kit was used as a template for PCR. The PCR reaction mixture consisted of approximately 1 μg cDNA obtained from brain, 10 nM of forward and reverse primers, 1.5 mM $MgCl_2$, 0.2 mM dNTP mix, 1.5U of Taq polymerase, and PCR buffer in 50 μL reaction volumes. Amplification of the targets was achieved by touchdown PCR with the following PCR cycle: 94°C for 2 min, 30 cycles of 95°C for 45 sec, 65°C for 30 sec, 72°C for 1 min, and final extension at 72°C for 10 min, which were repeated for 30 cycles. PCR reaction carried out with RNA of brain

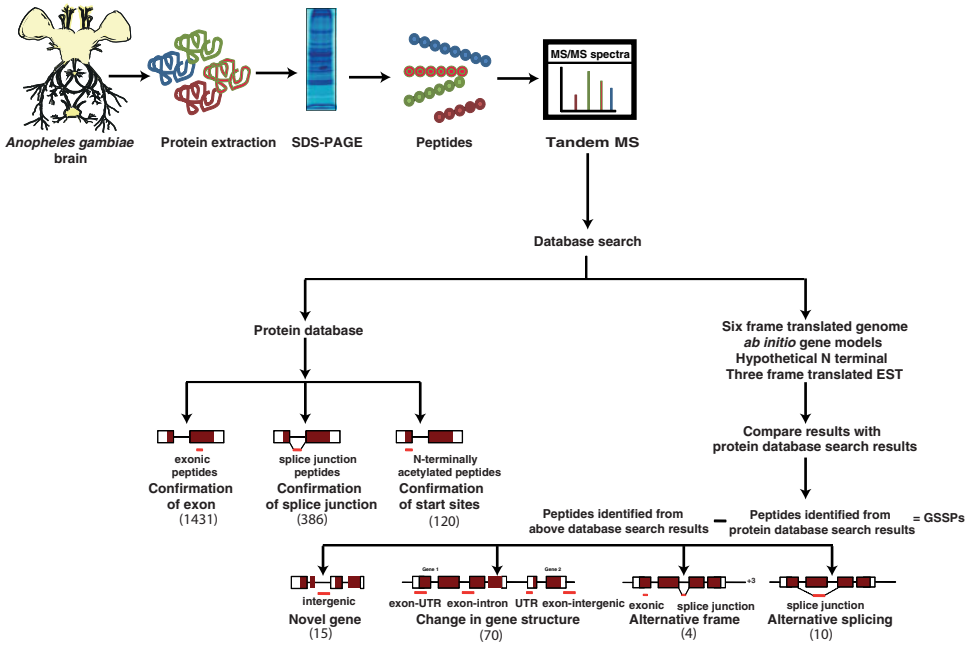


FIG. 1. Pictorial representation of the proteogenomic workflow. The peptides obtained from the protein database search were used to confirm the existence of the annotated Ensembl models, to experimentally confirm the known splice forms, and also to confirm the translational start sites of the annotated Ensembl models. GSSP classification led to identification of novel ORFs, and suggests various modifications in the Ensembl gene models by proposing gene joining events, novel exons, extension of exons, and also alternative splice forms. N-terminal acetylated peptides identified from Mascot protein database search were used to correct the annotated translational start sites or even to identify novel start sites.

served as negative control and to rule out genomic DNA contamination. Amplicon size was checked using DNA ladder on 2% agarose gel. Specific amplicons were purified by GeneJET gel extraction kit and subjected to sequencing by Sanger's method.

Results

Mass spectrometry data analysis

A total of 33 LC-MS/MS runs were carried out on the brain tissues of both *An. gambiae* male and female mosquitoes. Approximately 156,000 MS/MS spectra were acquired, out of which 79,401 were assigned to peptide sequences using Sequest and Mascot MS/MS search algorithms. Searches were submitted through Proteome Discoverer software (Thermo Scientific, version 1.3) to Sequest and Mascot (version 2.2) search engines with the parameters specified above. The total number of unique peptides identified from the protein search analysis was 9801 that corresponded to 1819 unique protein groups (Supplementary Table S1; supplementary material is available online at www.liebertpub.com/omi) from the analysis. Search against six frame translated genome database of *An. gambiae* PEST strain identified 6772 peptides. These peptides were compared with the protein database and those which mapped to the annotated protein coding genes were filtered out from the proteogenomic analysis. There were 101 peptides that did not map to any annotated protein coding genes in Ensembl and were designated genome search-specific peptides (hereafter referred to as GSSPs). These were the potential candidates for identifying novel events occurring in the genome. The novel events include all the new findings identified in the brain proteomic study, which has led to major revisions in the *An. gambiae* genome annotation. SNAP *ab intio* gene prediction database search overlapped 161 GSSPs, whereas 132 GSSPs showed an overlap with the three-frame translated EST sequences. In total, we identified 202 non-redundant GSSPs. Genome coordinates of these novel peptides were fetched using tblastn tool and only 183 GSSPs, which had a unique hit in the genome were retained for proteogenomic analysis. From these GSSPs, we could identify novel ORFs, splice variants and refine gene structures.

Confirmation of annotated Anopheles gambiae brain proteins

The strategy of searching MS/MS data against the genome and proteome of the studied organism to validate predicted ORFs and to identify novel ORFs has been described previously (Fermin et al., 2006; Kuster et al., 2001; Yates et al., 1995). In this study, high accuracy tandem mass spectrometry (MS/MS) data was queried against the proteome of *An. gambiae*. A total of 9801 (Supplementary Table S2) unique peptides were identified, which confirmed the protein coding potential of 1819 annotated *An. gambiae* genes. The peptides identified against known or predicted proteins were mapped to the genome and were categorized as exonic or splice junction peptides. The spliced peptides confirmed the respective exon boundaries, their frame of orientation, and their splice junction. One of the examples identified was in the protein coding gene AGAP001151-PA (14-3-3 protein) where 84% coverage was obtained with 24 unique peptides spanning all the 5 exons and the 4 splice boundaries of the gene model. We also obtained 65% coverage for the protein coding gene AGAP010895-PA where 99 unique peptides spanned 7 exons covering all the 6 splice boundaries.

The peptide sequences spanning exon–exon junctions provide evidence for confirmation of splice sites of predicted transcripts and novel splice variants. Peptides identified in the protein database search were mapped onto the transcript sequences from *Anopheles gambiae,* which also included splice isoforms to identify splice junction spanning peptides. We found a total of 553 unique peptides that spanned exon–exon junctions of Ensembl transcripts that supported 528 splice junctions from 386 genes. The list of exon–exon junctional peptides confirming the splicing events of annotated Ensembl gene models is provided in Supplementary Table S3.

Proteogenomic analysis

A potentially important aim of proteomics is to utilize the data in accurately defining the genome structure. This is achieved by identifying the prospective ORFs with their correct reading frame, their exon boundaries, enumerating their start sites, and determining their splice forms. Proteogenomic approach has been used by several groups in many microorganisms, such as *Toxoplasma gondii* (Xia et al., 2008), *Mycobacterium tuberculosis* (Kelkar et al., 2011), *Leishmania donovani* (Pawar et al., 2012), and *Candida glabrata* (Prasad et al., 2012). The computational steps involved in proteogenomic analysis has been described in a recent review by Renuse et al. (2011). The *An. gambiae* brain dataset was analyzed using different strategies of database searching and data analysis. A summary of all the novel findings identified in the proteogenomic analysis of *An. gambiae* has been provided in Table 1.

Confirmation and correction of start sites of known proteins

Accurate identification of protein start sites is a challenging task. Protein N-terminal peptides identified with the post-translational modification of acetylation were used for identifying such events. A total of 120 N-terminally acetylated peptides were identified from protein database search using Mascot search engine. These peptides provided confirmatory

Table 1. Summary of Novel Findings in Proteogenomic Analysis of *Anopheles gambiae*

Categories	*Novel identifications/ refinement of gene models*
Novel genes	15
N terminal extension of gene boundaries	10
C terminal extension of gene boundaries	11
Joining of gene	1
Gene modifications	21
Translated UTR	4
Correction of translational frame	4
Novel/alternate protein start site	10
Alternative splicing events	10

evidence for the start sites of 120 annotated Ensembl models. The peptide list has been provided in Supplementary Table S4. Hypothetical N-terminal database search using Mascot search engine identified 133 N-terminal acetylated peptides. On manual analysis and validation of the spectral assignments of these peptides, we assigned initiator methionines to eight annotated Ensembl models that have resulted in shortening or extension of gene models. Interestingly, these acetylated peptides could also assign initiator methionines in five Ensembl models that had no annotated start site. In addition, we identified two novel ORFs based on protein N-terminal acetylated peptide identification. The list of acetylated peptides, the type of gene structure modification along with the alternate gene coordinates are provided in Supplementary Table S5.

Intergenic peptides

Out of 146 novel GSSPs from our analysis, 80 peptides mapped outside annotated ORFs. These intergenic peptides were analyzed manually and the events were categorized as novel ORFs, N-terminal, or C-terminal extensions or joining of genes.

Identification of novel ORFs in intergenic regions. Manual analysis of 26 intergenic peptides led to the identification of 16 novel protein coding genes out of which one has recently been added in the new release AgamP3.7. Out of the 26 intergenic GSSPs, 7 were reported in the study by Chaerkady et al. (2011).

These novel protein coding regions were checked for conservation across species by using a protein blast tool. Out of the 15 novel ORFs identified, 14 were conserved across dipterans including *Aedes, Culex,* and *Drosophila*. Also, 4 of them were missed by the SNAP gene prediction program. The proteomics data furnish translation level evidence confirming the protein coding potential of the specific regions in the genome. Table 2 lists the details of the novel genes identified in this analysis with supporting peptide evidence. Twelve novel ORFs have been validated by RT-PCR and sequencing.

Several examples of potential novel genes were identified, a subset of which was also validated by cDNA sequencing. One such example had five peptides identified in the intergenic region between AGAP011106 and AGAP011109. On using alignment results from blast tool, this region was found to have 90% identity with RNA binding motif protein 4, also referred to as Lark protein *in Aedes aegypti* and *Drosophila melanogaster.* There have been knockdown studies on Lark protein, especially in Drosophila where it has been shown to exhibit pivotal roles in the circadian system affecting locomotory activity (Huang et al., 2009; Sundram et al., 2012). It plays a multifaceted role by targeting proteins encoding many developmental and physiological processes including neuronal survival, neurite growth and path finding, neuronal excitability, synaptic function, and in oogenesis (McNeil et al., 2009). The novel ORF was also identified in the study by Chaerkady et al. (2011), but no RT-PCR was done. Here, we have further validated the novel event by RT-PCR and sequencing. This novel gene, which was missed by Ensembl, could become one of the major candidates to be explored further for malaria control strategies.

Similarly we identified two intergenic peptides in the 2R chromosome where no genes were predicted in the vicinity of 4 Kb. SNAP gene prediction program did not predict any gene in this region. Other gene prediction programs used in the analysis were Fgenesh, Genemark, and Genscan. Among them, only Fgenesh predicted a 2 exon gene model in the genomic region. Further blastx gave additional evidence of homology with translocon-associated complex TRAP delta subunit in *Anopheles funestus*. This gene has been associated with receptor activity, component of endoplasmic reticulum and other membranes.

We identified another novel gene on the basis of two high scoring peptides. These peptides mapped within a 29 Kb region of the 2R chromosome where neither Ensembl nor SNAP predictions were available. However, Fgenesh predicted a single exon gene model in this region. Blastx showed alignment with a mitochondrial protein tricarboxylate transport protein in *Aedes aegypti*. Primers were designed for the single exon gene model and RT-PCR further confirmed the coding potential of the region. Three more novel genes were revealed based on peptides mapping in the regions where no Ensembl genes and no SNAP predictions were observed. However, deeper analyses using blast tool showed conservation with glutaminase protein in *Culex quinquefasciatus*, dystrobrevin in *Culex quinquefasciatus,* and disheveled-associated activator of morphogenesis 2 protein in *Culex quinquefasciatus* and *Aedes aegypti.* Primers were designed on the basis of Fgenesh prediction. Additionally, successful RT-PCR validation confirmed the presence of these novel genes. A representative figure of a novel gene identified has been shown in Supplementary Figure S1.

During the analysis of peptides hitting the introns of Ensembl gene models, we identified two ''gene within gene'' events. As represented in the Supplementary Figure S2, two peptides mapped to the 14 Kb long intron of AGAP012027 Ensembl model. Blastx analysis within 17 Kb genomic regions revealed a short ORF in the intronic region, which showed alignment with a hypothetical protein of 369 amino acids in *Anopheles darlingi.*

Identification of gene joining events. One of the major limitations in genome annotation using gene prediction methods is their inaccuracy in predicting gene structures by merging several genes as one single gene or sometimes mistakenly splitting a single true gene into several genes. This kind of observation was made during deep analysis of the brain dataset. Three examples were identified, where multiple peptides supported the occurrence of a single gene but Ensembl annotation predicted TWO separate gene models. The details of these unique examples are illustrated in Supplementary Table S6 and a representative figure is provided in Supplementary Figure S3. The figure shows two intergenic peptides obtained from our analysis that provided the evidence for the joining of the two Ensembl gene models AGAP011747 and AGAP011748. SNAP predicted the joining of the gene structure that also showed conservation with hypothetical protein in *Aedes aegypti.* One set of primers were designed between exon 2 of AGAP011747 and the exon 3 of SNAP model, and the other set was designed between exon 3 of AGAP011748 and the exon 5 of SNAP model. This example not only supported the joining of the two annotated gene models as a single gene, but it also provided evidence

of extension of the C- terminal boundary of the single gene model.

Another example is where two intergenic peptides were identified between AGAP010285 and AGAP010286. This is an example where SNAP did not predict a single gene by connecting the two gene models, but the Fgenesh program predicted a single gene model of 18 exons. This example was validated based on the Fgenesh gene model where three sets of primers were designed—one each across exon 1 and exon 3, exon 3 and exon 10, and exon 9 and exon 10 of the Fgenesh model. No homology evidence supporting the joining of gene was observed based on alignment results from blast tool. However, peptide evidence with additional level of RT-PCR validation provided the proof of existence of a single translated protein coding gene. This example has protein evidence in the current AgamP3.7 version and has now been annotated as a single gene model AGAP010286. Similarly another correction identified was the joining of the two Ensembl gene models AGAP011872 and AGAP011873. Multiple peptide identification from the brain dataset supported the chances of the split genes to be actually a single gene. This event has now been confirmed in the new release. The details of all the examples identified from the analysis have been provided in Supplementary Table S6.

Correction of predicted gene models

One of the requirements of a gene prediction tool is correct identification of the genomic coordinates of the coding exons in eukaryotic genomes. Gene annotation using existing computational approaches can often lead to erroneous predictions. A total of fourteen examples of C-terminal extensions and 21 N-terminal extensions of gene boundaries were identified in this study. The new release of AgamP3.7 reflects ten cases of N terminal extensions and three cases of C-terminal extension of gene boundaries that were annotated by us. The complete details of all the examples of gene refinement are provided in the Supplementary Table S6. Out of the 14 cases of C terminal extension of gene boundaries, 7 annotated genes did not end with a stop codon. The peptide data corrected the annotated gene boundaries by appropriate extension of the gene.

Out of the 21 events of N-terminal extensions, two representative examples are highlighted in Supplementary Figure S4A and S4B. The example shows 12 intergenic peptides, one peptide partially overlapping the N-terminal exon boundary and intergenic region, and one novel exon–exon junctional peptide mapped to the N-terminal region of the Ensembl model AGAP007763. The other is an example where three intergenic peptides supported the N-terminal extension of the AGAP009598 Ensembl model. The gene is further extended for 47 Kb. NCBI annotation is in agreement with our analysis for the protein voltage-gated calcium channel alpha2-delta subunit whose length is 1256 amino acids. However Ensembl has predicted an incomplete gene structure of 1124 amino acids. Two extension events were identified on the basis of multiple peptide evidence where the annotated Ensembl models AGAP008803 and AGAP010090 did not start with an initiator methionine.

The remaining 66 GSSPs have been broadly classified as intragenic peptides leading to different types of modifications in annotated Ensembl gene models. A separate category of ''Gene modification'' has been added for these kinds of examples that include novel exons, extension of exons and intron retentions. Twenty one such examples have been identified, among which three of them have been added in the new release. Figure 2 illustrates an example where a cluster of intergenic, UTR, exon–no gene junctional and novel exon–exon junctional peptides added a novel exon, extended a predicted exon and also the N and C-terminals of the Ensembl gene AGAP011700. A cumulative list of the identified GSSPs with the corresponding novel events is provided in Supplementary Table S6.

Evidence of translation in UTR regions

We found eight peptides that mapped to regions annotated as untranslated (UTR) in the 5′ or 3′ ends of five Ensembl gene models. Such events were identified on the basis of two types of peptides—peptides mapping to UTRs of predicted gene models, and junctional peptides spanning UTR and coding exon of a gene model. These peptides corrected the annotated N-terminal boundary of gene models by extension. Out of the five instances identified, two already show extension of coding region in the current release. One novel example identified from this study is illustrated in Supplementary Figure S5 where 2 peptides were identified GTEGSGLVSVLLEGPPNAGK (peptide hitting UTR) and VCSPDEMVGFSEGAK (spanning UTR-exon boundary) and provided the evidence that translation begins upstream of the annotated start methionine. The peptides also show conservation with Nsf1 protein in *Culex quinquefasciatus*. Hence, peptide evidence can clearly provide insight to the gene structure of an organism and correct the annotations of the predicted gene model.

Correction of reading frames

One of the biggest challenges in genome annotation is determination of the correct reading frame of translation for protein coding genes. Wrong assignments of reading frame can occur due to inaccurate prediction by gene finding programs or due to some biological events. These events could be attributed to either ribosomal slippage resulting in either −1 or +1 shift of reading frame or even due to skipping of a stretch of nucleotides (Farabaugh, 1996). Several cases of translational frameshifts have been reported in previous studies where two translational reading frames are expressed as a single protein. It is difficult to identify and even comprehend the exact reason behind the frame changing events. However, proteomics data can provide the definitive proof of the correct translational frame through peptide data. In this study, four such events were identified whose details are provided in Supplementary Table S6.

Figure 3A illustrates an example where the peptide VIVSYIPFYGGK mapped to a different reading frame from the predicted AGAP011484 protein which is 91 amino acids in length. There is a SNAP model predicted in the altered frame supporting the peptide. However, the SNAP predicted protein appears to be truncated to 54 amino acids due to an apparent frameshift. Blast analysis of the truncated protein showed alignment with ubiquinol cytochrome C oxidoreductase-subunit 6.4 kD-subunit, putative protein in *Aedes aegypti*. The annotated MS/MS spectrum of the peptide is also shown in the figure.

Table 2. List of Novel ORFs Identified from the Proteogenomic Analysis

Novel ORFs	*Peptide*	*Type of database search*	*# PSMs*	*Mascot score*	*Sequest score*	*Homology across species*	*Genome coordinates of Novel ORFs (According to Ensembl)*	*RT-PCR validation Yes/No*	*Identified in previous study (Chaerkady et al)*
Novel ORF 1	TQHEQQQSSTVAPGNP VGRDEAINSK	Six frame genome	1	–	6.16	Conservation with hypothetical protein in *Culex quinquefasciatus*	Chr2L:27324952-27329417	Yes	No
	LTETEALEELK	SNAP protein	1	–	3.83				No
	LLYTQLQR	Six frame genome	1	58	–				No
	ESVEQGGGGGcGG VTEPPAR	SNAP protein	1	–	3.48				No
Novel ORF 2	FGSVVEcDIVR	SNAP protein	1	–	3.65	Conservation with RNA binding motif protein 4,lark in *Aedes aegypti*	Chr3L:17006847-17008327	Yes	No
	TQPSELRPLFEK	SNAP protein	1	–	3.62				Yes
	NYGFVHLDPTGDVNEAIR	SNAP protein	1	–	3.75				Yes
	YGTVVECDVVK	Six frame genome	1	63	–				No
	ASSSSYEAIFSR	SNAP protein	1	–	3.04				No
Novel ORF 3	AGEDVQSVKPLTTVT VNFPGAYK	EST three frame	2	–	7.26	Conservation with translocon-associated complex TRAP delta subunit in *Anopheles funestus*	Chr2R_hap_34: 36778767-36779372	Yes	No
	LFDEESYAAVR	EST three frame	1	–	3.72				No
Novel ORF 4	DMYKGDDPSKPVPK	EST three frame	1	–	3.88	Conservation with tricarboxylate transport protein, mitochondrial in *Culex sps*	Chr2R_hap_33: 36134233-36135498	Yes	No
	YKNTADcALQIWK	EST three frame	1	–	4.76				No
Novel ORF 5	SGPNAPLPQPLVAFK	SNAP protein	1	–	4.23	Conservation with conserved hypothetical protein in *Culex sps*	Chr3R:18659172-18660604	Yes	No
Novel ORF 6	LASSLLQQQQGSSK	SNAP protein	1	–	3.75	No conservation across species	Chr3R:18659172-18660604	Yes	Yes
Novel ORF 7	AEQAAGLDEEAAVTGK	EST three frame	4	90	5.27	Conservation with hypothetical protein AND_09996 in *Anopheles darlingi*	Chr3L:25648000-25652718	No	Yes
Novel ORF 8	LGAQVVHQYIGQEPSGR	EST three frame	3	–	6.73	Conservation with glutaminase protein in *Culex quinquefasciatus*	Chr2R_hap_34: 36731693-36735313	Yes	No
Novel ORF 9	TPGGSAPDPVQIGLDSSR	EST three frame	1	–	5.45	Conservation with dystrobrevin in *Culex quinquefasciatus*	Chr2R_hap_34: 36766757-36775522	Yes	No
Novel ORF 10	LQDPDVAPLAIDVR	EST three frame	2	–	4.3	Conservation with disheveled-associated activator of morphogenesis 2 in *Culexsps*	Chr2R_hap_31: 35222948-35228922	Yes	No

(*continued*)

TABLE 2. (CONTINUED)

Novel ORFs	*Peptide*	*Type of database search*	*# PSMs*	*Mascot score*	*Sequest score*	*Homology across species*	*Genome coordinates of Novel ORFs (According to Ensembl)*	*RT-PCR validation Yes/No*	*Identified in previous study (Chaerkady et al)*
Novel ORF 11	ENFLNPTAVVDLK	Six frame genome	4	68	–	Conservation with hypothetical protein AND_11499 in *Anopheles darlingi, EFR24133*	Chr3R:36614226-36618384	Yes	No
Novel ORF 12	MEAVAKPSSTR	Hypothetical N terminal	2	61	–	Conservation with conserved hypothetical protein in *Culex quinquefasciatus* XP_001850371.1	Chr3R:37221557-37223073	No	No
Novel ORF 13 (Gene within gene)	YSLGVGEGETNLPDGK IFPEAETVADSA LPYTAFR	SNAP protein Six frame genome	1 1	– 52	3.48	Conservation with hypothetical protein in *Anopheles darlingi*	Chr3L:36669088-36670194	Yes	Yes No
Novel ORF 14 (Gene within gene)	HVELYNVGNTK	SNAP protein	1	–	3.67	Conservation with Dbuz\CG31363-PC *Drosophila buzzatii*	Chr2R:3921755-3922718	Yes	Yes
Novel ORF 15	AIVNAGDFPADVK	Six frame translated	1	–	4.14	No conservation across species	ChrUNKN:5878711-5884354	No	Yes
Novel ORF 16	AAQTDISALAQQWR	SNAP protein	1	–	3.72	Conservation with conserved hypothetical protein *Culex quinquefasciatus*	Chr3L:20805855-20806178	No	No

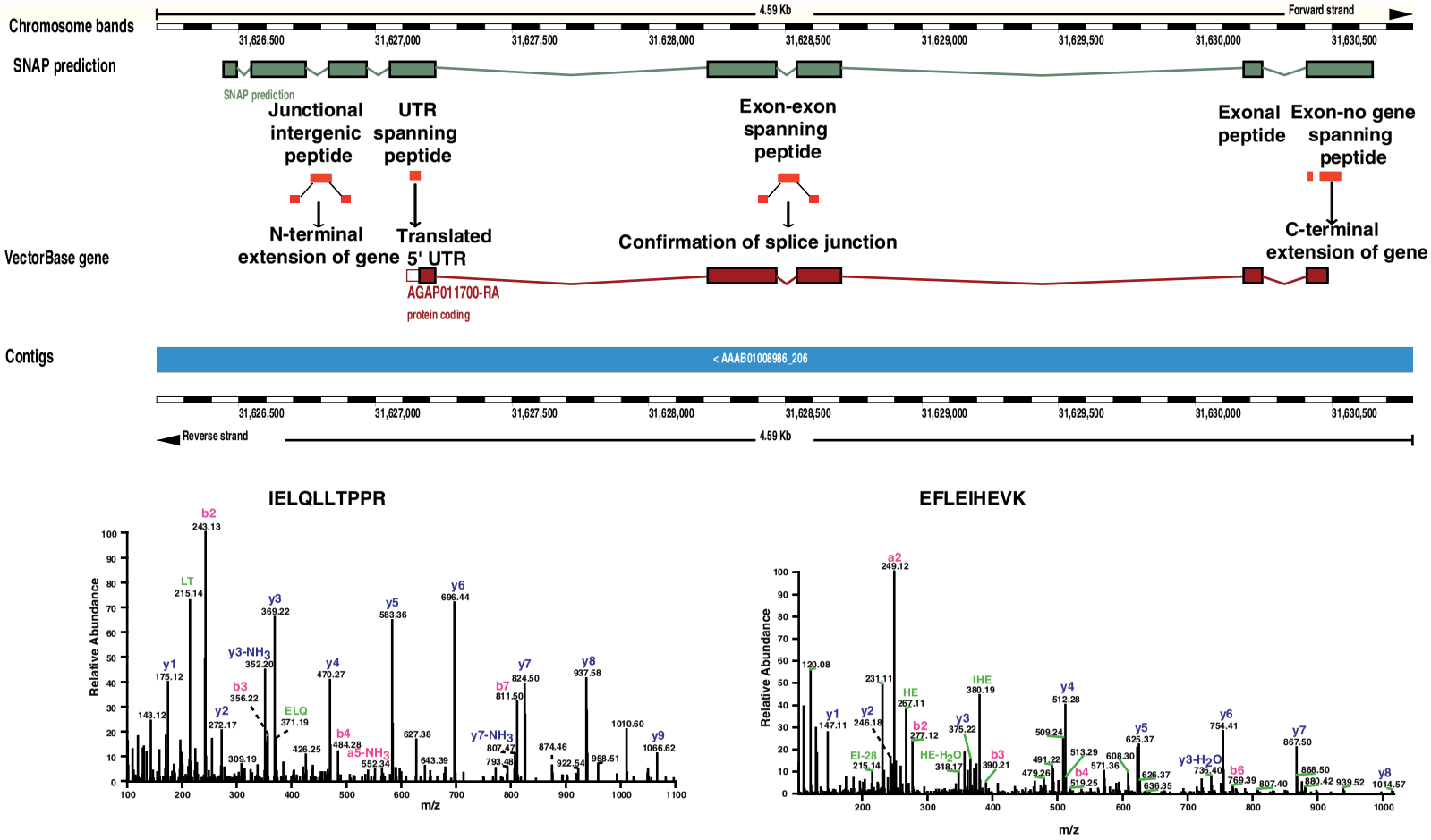


FIG. 2. Correction of the gene structure of the annotated Ensembl models. An example of ‘‘gene modification’’ where multiple peptides belonging to different categories of peptides as specified in the workflow have modified the entire structure of the gene model AGAP011700 in different ways. Two representative MS/MS spectra supporting the identification are provided.

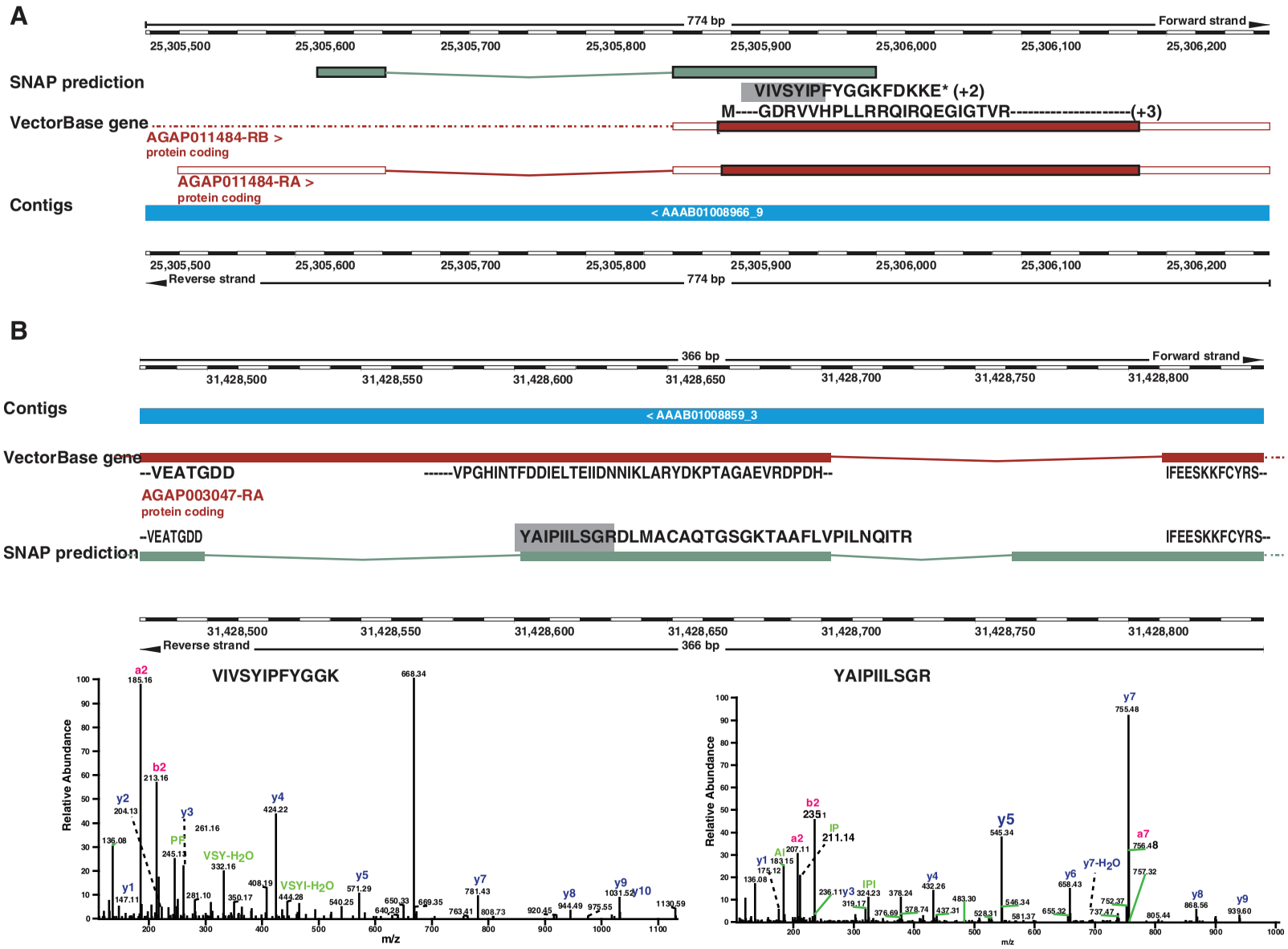


FIG. 3. Correction of frame of the Ensembl models. **(A)** The coordinates of a single peptide VIVSYIPFYGGK fell within the coordinates of an exon of the Ensembl model AGAP011484 (+3) but was translated in a different reading frame (+2) from the annotated model. This led to the identification of a truncated ORF of 54 amino acids with an in frame stop codon. The truncated protein showed 100% identity with ubiquinol cytochrome C oxidoreductase-subunit 6.4kD-subunit, putative protein in *Aedes aegypti*. **(B)** An example of both frame change and of a novel splice form. The peptide was supported by a SNAP model representing a splice form that differed from the predicted model, with the second exon splitting and frame changing in the middle part of the protein coding exon. The first exon of the SNAP model and a part of the second exon of the AGAP003047 model were in the same translational reading frame.

The example shown in Figure 3B is a peptide YAIPIILSGR, which mapped to a predicted *An. gambiae* protein AGAP003047 that has four transcripts but was located in a translational reading frame different from the predicted polypeptide. However, the peptide showed 100% identity to the SNAP model. On further analysis, it was found that the protein product was partly encoded in the normal frame upstream of the shift and partly in the shifted frame downstream of it and then again translated back in the original reading frame. The reason behind the frame change can be associated with alternative splicing, and therefore this example can also be cited as a novel splice form. Also, the frame shift peptide showed conservation with hypothetical protein in *Anopheles darlingi.*

Alternative splicing events

Alternative splicing is one of the major mechanisms in eukaryotes contributing to its protein diversity. Identifying this mechanism accurately is one of the biggest challenges in the genome annotation process. In this study, we identified 15 novel spliced peptides from the SNAP prediction database and EST three frame translated database searches. These spliced peptides were mapped back to the genome to show the exact location of split across exons. cDNA sequencing gave proof of evidence for use of non-canonical splice site sequences in two cases. The rest of them were GT-AG splice rule compliant. Non-canonical splice site prediction cannot be taken care of by most of the available splice predictor or gene prediction algorithms. This is evident from our dataset where our gene models based on peptide data and cDNA sequencing differed from the models predicted by Ensembl, as well as the gene prediction programs. The details of the novel junctional peptides, splicing events, the annotated intron coordinates, and the alternate intron coordinates, with their splice site sequences are provided in Supplementary Table S7. The brain dataset revealed two cases of alternative donors, three of alternative acceptors, two of alternative sites, and four cases of novel splice forms. Three representative examples of alternative splice form types as described in the workflow have been shown in Supplementary Figures S6 and S7.

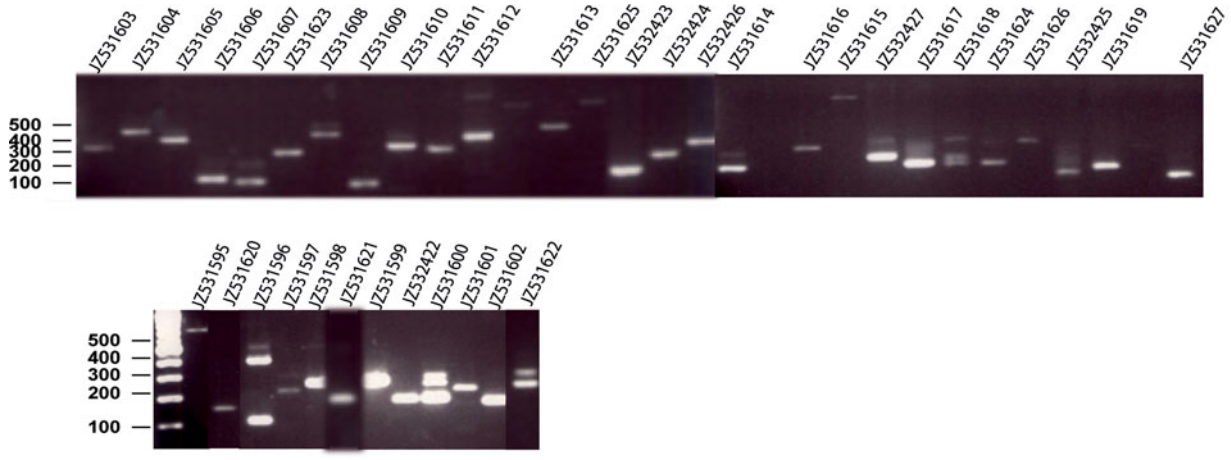
Validation of mass-spectrometry derived data by RT-PCR validation

RT-PCR validation was done for a subset of the novel events identified. These included 14 novel genes (out of which two have been identified in the new AgamP3.7 release), nine alternative splicing events, three gene joining events, two gene within gene, and 11 other gene modification events. The cDNA sequences have been submitted to NCBI Genbank. The PCR gel image of novel events with their respective GenBank accession numbers are shown in Figure 4.

Discussion

Overview of identified proteins

Biomart version 0.7 was used to fetch gene ontology terms for the proteins identified from the study (Kinsella et al., 2011). GO tool assigned biological processes to 1546 protein coding genes. The remaining 273 were classified as an unannotated class of proteins. Supplementary Table S1 (PXD000630) represents the list of identified proteins. The protein coding genes identified were found to be involved in



JZ531595 - Novel ORF
JZ531596 - Novel ORF
JZ531597 - Novel ORF
JZ531598 - Novel ORF
JZ531599 - Novel ORF
JZ531600 - Novel ORF
JZ531601 - Novel ORF
JZ531602 - Novel ORF
JZ531603 - Gene extension
JZ531604- Gene extension
JZ531605- Gene extension
JZ531606- Gene extension
JZ531607- Gene extension
JZ531608- Gene extension
JZ531609- Gene extension
JZ531610- Gene extension
JZ531611- Gene fusion
JZ531612- Gene fusion
JZ531613- Gene fusion
JZ531614 - Gene modification
JZ531615 - Gene modification
JZ531616 - Gene modification
JZ531617 - Alternative splicing
JZ531618 - Alternative splicing
JZ531619 - Alternative splicing
JZ531620 - Novel ORF
JZ531621 - Novel ORF
JZ531622 - Gene extension
JZ531623 - Gene extension
JZ531624- Alternative splicing
JZ531625- Gene fusion
JZ531626 - Alternative splicing
JZ531627 - Alternative splicing
JZ532422 - Novel ORF
JZ532423- Gene fusion
JZ532424- Gene fusion
JZ532425 - Alternative splicing
JZ532426- Gene fusion
JZ532427 - Alternative splicing

FIG. 4. Gel image of RT-PCR validation done on a subset of the identified novel events. The two panels of gel images represent the examples of novel ORFs, extension of gene models, gene joining events, gene within gene, and gene modification, which were validated by RT-PCR and cDNA sequencing. The Genbank accessions are provided for each example validated.

various biological processes such as transport, transcription/translation, metabolism, neurotransmission, olfaction, and signal transduction pathways. Figure 5 shows the distribution of the identified proteins based on their biological processes. Major protein functions represented in the brain samples of the vector included proteins involved in energy production and conversion by major metabolic processes like glycolysis, TCA cycle, oxidative phosphorylation, amino acid synthesis, and signal transduction mechanisms.

Based on spectral counts to estimate the relative abundance, the top 25 most abundant proteins identified in this study, show enrichment of metabolic, oxidative phosphorylation, phagosome, amino acid synthesis, and hippo signaling pathways related proteins. Proteins involved in metabolic processes represented the largest fraction in the group. Examples such as sodium and potassium ion transporting ATPase subunits alpha, beta and gamma, also referred to as the Na pump, are among the topmost abundant proteins identified in the brain study. The Na/K ion ATPase is an ion pump that translocates sodium and potassium ions across plasma membrane, maintaining the homeostatic balance in the cells. The protein has been extensively studied in various organisms such as humans and drosophila, highlighting its importance in maintaining normal neuronal functions in an organism. Mutational studies and targeted knockdown studies have revealed that perturbations in the Na/K ATPase can lead to severe dysfunction in behavior, neuronal excitability, and defects in auditory mechanosensation (Roy et al., 2013), visual and motor sensory system (Palladino et al., 2003). Molecular characterization of the protein in gambiae is yet to be described. Elevated levels of this kind of protein reflect more active protein synthesis in the brain of *An. gambiae* to develop the neuronal machinery in the vector. Molecular analysis of this molecule in *An. gambiae* can provide a valuable model for elucidating the neuronal mechanism which can affect the longevity of the vector.

In addition, proteins related to chromatin modeling, vesicular transport, cytoskeletal and signaling processes were also abundant. Nuclear proteins play a central role in developing the transcriptional machinery. Several histone families of proteins H4, H2A, H2B.1, and H2B.2 were detected in abundance in the brain sample. One of the abundant proteins AGAP012871-PA was unannotated. Blast analysis showed 99% identity with H2B protein in *Drosophila grimshawi*. Other highly expressed proteins were cytoskeletal proteins such as tubulin beta, tubulin alpha 4A, and vesicle transport proteins such as synaptic vesicle-associated integral membrane protein, which may be associated with synaptic neuronal transmission. Neural proteins like Armadillo segment polarity protein (Arm) identified in the study have been extensively studied in Drosophila. It plays a pivotal role in transducing Wingless (Wg) signal in the neural cells (Loureiro et al., 1998). The wingless gene is associated with generating diversity in neuroblasts in CNS development (Chu-LaGraff et al., 1993). Each neuroblast eventually divides to form neurons expressing different sensory and motor receptors, neurotransmitters, and other cell surface molecules. Thus, the segment polarity class of genes serves as potential candidates for studying the diversity pattern of neuroblasts in central nervous system development of *An. gambiae*.

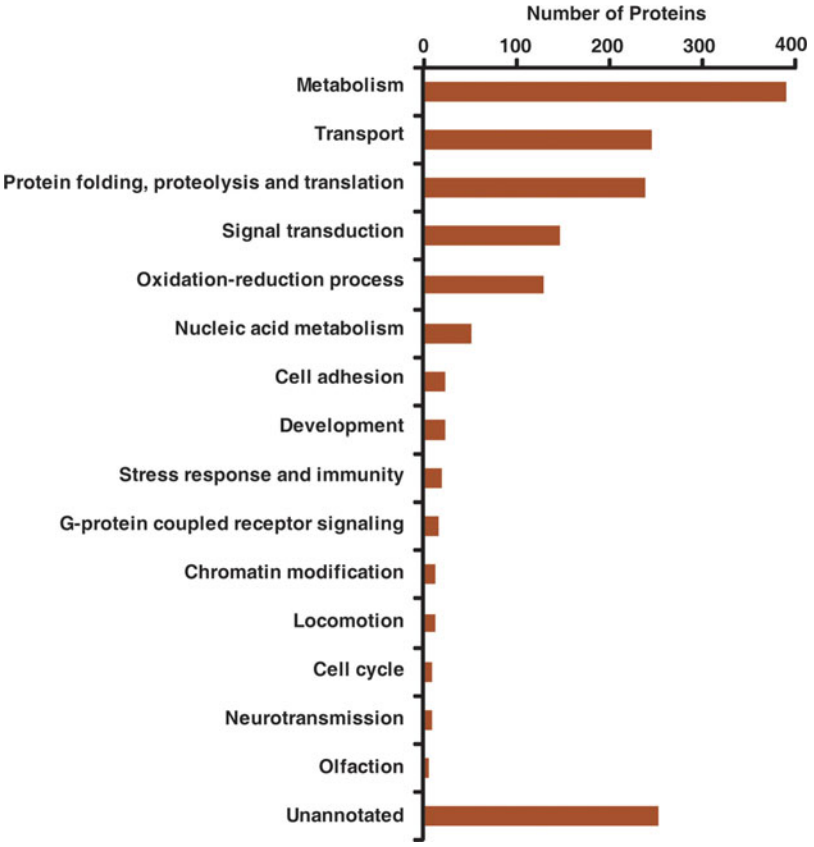


FIG. 5. Functional categorization of the identified proteins based on Gene ontology annotation. The distribution of the identified proteins was based on biological processes. 391 proteins were involved in metabolism. 241 proteins were involved in the translational machinery. 273 proteins were not assigned any GO terms and have been put under an unannotated class. A small subset of the identified proteins was involved in olfaction, neurotransmission, stress response, and development.

Enrichment of GPCRs in the brain study

Proteins involved in the communication system of mosquitoes can serve as promising targets which directly impact the vectorial capacity of *An. gambiae*. G-protein coupled receptors are the class of proteins mediating sensory functions such as photo and chemo receptions, synaptic processes, and hormonal regulations (Hill et al., 2002). A total of 276 GPCRs have been identified in the *An. gambiae* genome using computational approaches. The *An. gambiae* GPCRs have been majorly categorized into five classes, out of which we identified 10 proteins belonging to four classes, namely rhod(opsin) receptor, secretin receptor, metabotropic glutamate receptor, and atypical 7TM proteins family. Two amine receptors GPRmtn (melatonin) and GPRtyr (tyramin) proteins and one rhodopsin receptor were identified in *An. gambiae* through homology based bioinformatic approaches. The brain proteomics data provide the experimental evidence of its existence. Further functional studies of these proteins can provide an insight into the phototransduction cascade in the vector, which influence the mosquito behavior. Other vision-related genes identified in the study, for example, eye-specific G protein β-subunit gene, cyclophillin, and photoreceptor-specific cytidine diphosphate diacylglycerol synthase, show a rhythmic pattern of expression in presence of light and darkness. The regulation of these proteins influences the vector's locomotory behavior and ultimately plays an important role in host localization (Rund et al., 2011). Similarly, proteomic evidence related to the olfactory pathway like odorant binding protein 9 and phospholipase C-β have been found in the brain study. Previous studies have shown that OBP9 is widely expressed in male antennae, hemolymph, and even in pre-adult stages (Mastrobuoni et al., 2013). Now, with proteomics data we have confirmed its existence in the brain too. Among the olfactory repertoire, OBP9 is reported to be the most abundant and that could be one of the major reasons of other olfactory proteins not being detected in our dataset. Phospholipase C-β is one of the key effector odorant enzymes of the inositol 1, 4, 5-trisphosphate (IP3) signal transduction pathway, which gets activated when an odorant receptor protein directly interacts with an odorant, forming a complex with G protein. The activated enzymes would then release secondary messengers affecting the neuronal membrane potential (Zwiebel et al., 2004). Studies in Drosophila have shown that mutation in phospholipase C-β reduced odorant responses to chemical stimuli. A more detailed characterization and knowledge of the signaling pathways associated with olfaction can prove to be highly beneficial to identify new molecules that can be used to develop new strategies like better repellants for transmission control.

The other categories of GPCRs enriched in our study are the secretin-like (class B) family of G protein-coupled receptors (GPCRs) and Class C metabotropic glutamate receptor. Class B receptors majorly interact with large, glycoprotein hormones such as secretin, glucagons, and parathyroid hormone regulating neuronal survival. Out of the three Class B family receptors found in the study, two are orphan or putative GPCR receptors GPorphb2, GPorphb1, which need functional characterization. Latrophilin receptors, a neuronal adhesion G-protein-coupled receptor, was the other class B GPCR identified in the study. Previous studies in nematodes have shown that it controls neurotransmitter release on binding with specific ligands in adult brains (Silva et al., 2010). In-depth study of latrophilin-related GPCR signaling pathway in *An. gambiae* is required to understand their ligand specific functionalities. A number of proteins were also involved in the development and regulation of the cellular machinery like histone proteins and translation initiation factors. Studies have shown that these proteins may also be involved in the circadian regulation of different biological activities such as metabolism, transcription and translation machinery and show rhythmic pattern of expression.

An. gambiae is one of the better studied malaria vectors with its genome re-annotated in several attempts by multiple groups. In spite of this, the genome annotation is far from being complete. This is demonstrated by the fact that we could identify several novel findings that were missed by the current Ensembl annotation pipeline. The major advantage of a proteomics-based annotation system is that the peptide data provide the direct evidence of translation of a genomic region. The 15 novel protein coding genes identified and validated in the brain study have shown inference to be involved in major neural processes such as neuronal activities, locomotion, and morphogenesis, to name a few. The description has been provided in the above respective sections. Peptide data can ascertain other very critical aspects in a protein coding gene such as the translational start and stop of a gene and even the correct prediction of the reading frame. The correct annotation of the ATG start site is always difficult using cDNA information or even homology based approach. Using mass spectrometry derived N-terminal acetylated peptide data, the direct evidence of the initiator methionine can be established. As identified in our study, we could assign or correct start sites for ten Ensembl gene models.

Another important application of proteogenomically identified peptide is identification of alternative splice forms. Alternative splicing can use exons in different reading frames of translation. This accounts for the diversity of the proteome. In the current dataset, three such novel splice forms were identified. Figure 6A represents a type of novel splice form, which arose not only due to alternative splicing mechanism but also due to the partial change in the translational reading frame of one of the two exons spanning the junctional peptide in the annotated Ensembl model. In the brain dataset, one of the major highlights has been the identification of a novel splice form in an unsequenced stretch of the *Anopheles gambiae* genome as described below.

Identification of a novel splice form in an unsequenced region of Anopheles gambiae *genome*

Sequence based annotation efforts can have multidimensional applications not only in amending genome annotation but also in identifying sequencing flaws. Figure 6B depicts an example where three peptide sequences QYRPVVYSNTIQSLVAILR, IIHESGFTSEDFK, and AMPNLSIAFGNNEREcDAK were identified from three frame translated EST database search. Blast tool gave no hit against the *Anopheles gambiae* genome. But NCBI protein blast gave a 100% hit against G protein alpha subunits AgOn and AgOa in *An. gambiae* (AGAP005773). Both the transcripts contain stretches of sequences that were not sequenced in *An. gambiae* genome project (Holt et al., 2002; Lawniczak et al., 2010).

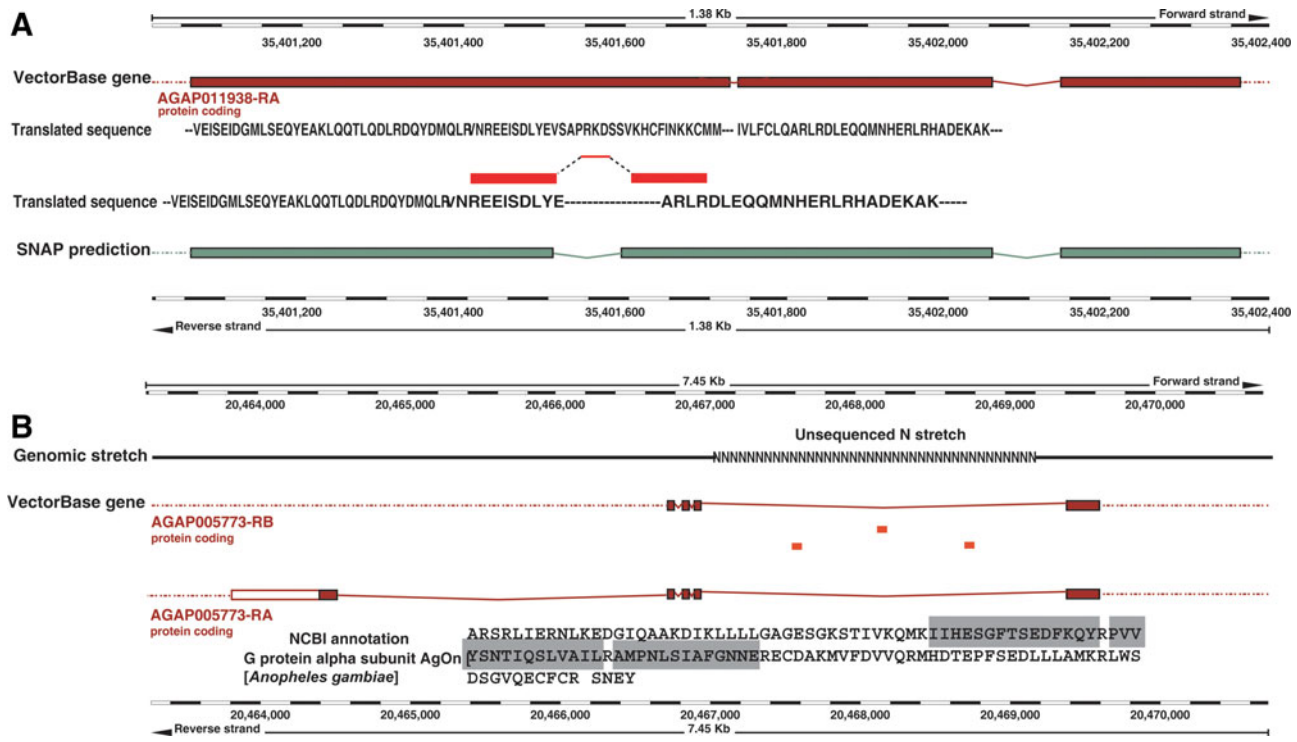


FIG. 6. Novel splice forms. **(A)** A novel splice junction peptide VNREEISDLYEAR represents an example of a novel splice form arising due to splicing in exons translated in different reading frames. **(B)** is an example of a novel splice peptide AMPNLSIAFGNNERECDAK identified in an unsequenced stretch of *Anopheles gambiae* genome. The NCBI annotation identifies the novel splice form as a GTP binding protein.

The identified peptide lies within the unsequenced stretch of the genome (~2Kb) between exon 4 and exon 5 of transcript AGAP005773-RA and exon5 and exon 6 of transcript AGAP005773-RB. The transcripts derived from the identified Gα-genes have been confirmed and characterized by RT-PCR and have been shown to be localized in axon bundles and at the base of sensilla chaetica, which is believed to function in mosquito mechanosensation (Rutzler et al., 2006). The study conducted by the Rutzler group did not find any amplification in adult tissues but only in embryonic stages; however the high-quality proteomic data confirmed the existence of the G protein in adult brain tissues also.

Conclusions

The aim of this study was to provide a catalog of the proteins identified in the brain of *An. gambiae* mosquitoes based on experimental evidence and to additionally use the data to refine its genome annotation. We report here the identification of a set of novel protein coding regions in the *An. gambiae* genome obtained by high resolution tandem mass-spectrometry. With the proteogenomic approach, we were able to obtain a more comprehensive and accurate catalog of protein-coding genes in the *An. gambiae* genome. The study has also identified some novel proteins that could be studied further as targets for malaria control strategies.

Acknowledgments

We thank the Department of Biotechnology (DBT), Government of India, for research support to the Institute of Bioinformatics and the pilot project award from the Johns Hopkins Malaria Research Institute. T. S. Keshava Prasad is a recipient of the research grant on ''Development of Infrastructure and a Computational Framework for Analysis of Proteomic Data'' from DBT (BT/01/COE/08/05). Sutopa B. Dwivedi, Babylakshmi Muthusamy, Raja Sekhar Nirujogi are recipients of Senior Research Fellowships from Council of Scientific and Industrial Research (CSIR), India. Gourav Dey is a recipient of Junior Research Fellowship from University Grants Commission (UGC) Government of India. Harsha Gowda is a Wellcome Trust/DBT India Alliance Early Career Fellow.

Author Disclosure Statement

The authors declare no conflict of interest. No competing financial interests exist.

References

Amenya DA, Chou W, Li J, et al. (2010). Proteomics reveals novel components of the Anopheles gambiae eggshell. J Insect Physiol 56, 1414–1419.

Carey AF, Wang G, Su CY, Zwiebel LJ, and Carlson JR. (2010). Odorant reception in the malaria mosquito Anopheles gambiae. Nature 464, 66–71.

Castellana N, and Bafna V. (2010). Proteogenomics to discover the full coding content of genomes: A computational perspective. J Proteomics 73, 2124–2135.

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.

Chu-LaGraff Q, and Doe CQ. (1993). Neuroblast specification and formation regulated by wingless in the Drosophila CNS. Science 261, 1594–1597.

Dani FR, Francese S, Mastrobuoni G, et al. (2008). Exploring proteins in Anopheles gambiae male and female antennae through MALDI mass spectrometry profiling. PLoS One 3, e2822.

Dinglasan RR, Devenport M, Florens L, et al. (2009). The Anopheles gambiae adult midgut peritrophic matrix proteome. Insect Biochem Mol Biol 39, 125–134.

Farabaugh PJ. (1996). Programmed translational frameshifting. Annu Rev Genet 30, 507–528.

Fermin D, Allen BB, Blackwell TW, et al. (2006). Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. Genome Biol 7, R35.

Harsha HC, Molina H, and Pandey A. (2008). Quantitative proteomics using stable isotope labeling with amino acids in cell culture. Nat Protoc 3, 505–516.

He N, Botelho JM, McNall RJ, et al. (2007). Proteomic analysis of cast cuticles from Anopheles gambiae by tandem mass spectrometry. Insect Biochem Mol Biol 37, 135–146.

Hernandez C, Waridel P, and Quadroni M. (2014). Database construction and peptide identification strategies for proteogenomic studies on sequenced genomes. Curr Top Med Chem 14, 425–434.

Hernandez LG, Lu B, da Cruz GC, et al. (2012). Worker honeybee brain proteome. J Proteome Res 11, 1485–1493.

Hill CA, Fox AN, Pitts RJ, et al. (2002). G protein-coupled receptors in Anopheles gambiae. Science 298, 176–178.

Holt RA, Subramanian GM, Halpern A, et al. (2002). The genome sequence of the malaria mosquito Anopheles gambiae. Science 298, 129–149.

Huang Y, Howlett E, Stern M, and Jackson FR. (2009). Altered LARK expression perturbs development and physiology of the Drosophila PDF clock neurons. Mol Cell Neurosci 41, 196–205.

Hummon AB, Richmond TA, Verleyen P, et al. (2006). From the genome to the proteome: Uncovering peptides in the Apis brain. Science 314, 647–649.

Huybrechts J, Bonhomme J, Minoli S, et al. (2010). Neuropeptide and neurohormone precursors in the pea aphid, Acyrthosiphon pisum. Insect Mol Biol 19, 87–95.

Kalume DE, Okulate M, Zhong J, et al. (2005a). A proteomic analysis of salivary glands of female Anopheles gambiae mosquito. Proteomics 5, 3765–3777.

Kalume DE, Peri S, Reddy R, et al. (2005b). Genome annotation of Anopheles gambiae using mass spectrometry-derived data. BMC Genomics 6, 128.

Kelkar DS, Kumar D, Kumar P, et al. (2011). Proteogenomic analysis of Mycobacterium tuberculosis by high resolution mass spectrometry. Mol Cell Proteomics 10, M111 011627.

Kinsella RJ, Kahari A, Haider S, et al. (2011). Ensembl BioMarts: A hub for data retrieval across taxonomic space. Database (Oxford) 2011, bar030.

Krug K, Nahnsen S, and Macek B. (2011). Mass spectrometry at the interface of proteomics and genomics. Mol Biosyst 7, 284–291.

Kuster B, Mortensen P, Andersen JS, and Mann M. (2001). Mass spectrometry allows direct identification of proteins in large genomes. Proteomics 1, 641–650.

Lawniczak MK, Emrich SJ, Holloway AK, et al. (2010). Widespread divergence between incipient Anopheles gambiae species revealed by whole genome sequences. Science 330, 512–514.

Li AQ, Popova-Butler A, Dean DH, and Denlinger DL. (2007). Proteomics of the flesh fly brain reveals an abundance of upregulated heat shock proteins during pupal diapause. J Insect Physiol 53, 385–391.

Li J, Hosseini Moghaddam SH, Chen X, Chen M, and Zhong B. (2010). Shotgun strategy-based proteome profiling analysis on the head of silkworm Bombyx mori. Amino Acids 39, 751–761.

Li JY, Chen X, Fan W, et al. (2009). Proteomic and bioinformatic analysis on endocrine organs of domesticated silkworm, Bombyx mori L. for a comprehensive understanding of their roles and relations. J Proteome Res 8, 2620–2632.

Lin MF, Carlson JW, Crosby MA, et al. (2007). Revisiting the protein-coding gene catalog of Drosophila melanogaster using 12 fly genomes. Genome Res 17, 1823–1836.

Liu C, Pitts RJ, Bohbot JD, Jones PL, Wang G, and Zwiebel LJ. (2010). Distinct olfactory signaling mechanisms in the malaria vector mosquito Anopheles gambiae. PLoS Biol 8, e1000467.

Loureiro J, and Peifer M. (1998). Roles of Armadillo, a Drosophila catenin, during central nervous system development. Curr Biol 8, 622–632.

Mastrobuoni G, Qiao H, Iovinella I, et al. (2013). A proteomic investigation of soluble olfactory proteins in Anopheles gambiae. PLoS One 8, e75162.

McNeil GP, Kaur M, Purrier S, and Kang R. (2009). The Drosophila RNA-binding protein Lark is required for localization of Dmoesin to the oocyte cortex during oogenesis. Dev Genes Evol 219, 11–19.

Okulate MA, Kalume DE, Reddy R, et al. (2007). Identification and molecular characterization of a novel protein Saglin as a target of monoclonal antibodies affecting salivary gland infectivity of Plasmodium sporozoites. Insect Mol Biol 16, 711–722.

Olsen JV, de Godoy LM, Li G, et al. (2005). Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. Mol Cell Proteomics 4, 2010–2021.

Palladino MJ, Bower JE, Kreber R, and Ganetzky B. (2003). Neural dysfunction and neurodegeneration in Drosophila Na+/K+ ATPase alpha subunit mutants. J Neurosci 23, 1276–1286.

Pandey A, and Mann M. (2000). Proteomics to study genes and genomes. Nature 405, 837–846.

Paskewitz SM, and Shi L. (2005). The hemolymph proteome of Anopheles gambiae. Insect Biochem Mol Biol 35, 815–824.

Pavlides SC, Pavlides SA, and Tammariello SP. (2011). Proteomic and phosphoproteomic profiling during diapause entrance in the flesh fly, Sarcophaga crassipalpis. J Insect Physiol 57, 635–644.

Pawar H, Sahasrabuddhe NA, Renuse S, et al. (2012). A proteogenomic approach to map the proteome of an unsequenced pathogen—Leishmania donovani. Proteomics 12, 832–844.

Prasad TS, Harsha HC, Keerthikumar S, et al. (2012). Proteogenomic analysis of Candida glabrata using high resolution mass spectrometry. J Proteome Res 11, 247–260.

Predel R, Neupert S, Garczynski SF, et al. (2010). Neuropeptidomics of the mosquito Aedes aegypti. J Proteome Res 9, 2006–2015.

Renuse S, Chaerkady R, and Pandey A. (2011). Proteogenomics. Proteomics 11, 620–630.

Roy M, Sivan-Loukianova E, and Eberl DF. (2013). Cell-type-specific roles of Na+/K+ ATPase subunits in Drosophila auditory mechanosensation. Proc Natl Acad Sci USA 110, 181–186.

Rozen S, and Skaletsky H. (2000). Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132, 365–386.

Rutzler M, Lu T, and Zwiebel LJ. (2006). Galpha encoding gene family of the malaria vector mosquito Anopheles gambiae: Expression analysis and immunolocalization of AGalphaq and AGalphao in female antennae. J Comp Neurol 499, 533–545.

Silva JP, and Ushkaryov YA. (2010). The latrophilins, ''split-personality'' receptors. Adv Exp Med Biol 706, 59–75.

Sundram V, Ng FS, Roberts MA, Millan C, Ewer J, and Jackson FR. (2012). Cellular requirements for LARK in the Drosophila circadian system. J Biol Rhythms 27, 183–195.

Uno Y, Fujiyuki T, Morioka M, Takeuchi H, and Kubo T. (2007). Identification of proteins whose expression is up- or down-regulated in the mushroom bodies in the honeybee brain using proteomics. FEBS Lett 581, 97–101.

Vizcaino JA, Cote RG, Csordas A, et al. (2013). The PRoteomics IDEntifications (PRIDE) database and associated tools: Status in 2013. Nucleic Acids Res 41, D1063–1069.

Wolschin F, Munch D, and Amdam GV. (2009). Structural and proteomic analyses reveal regional brain differences during honeybee aging. J Exp Biol 212, 4027–4032.

Xia D, Sanderson SJ, Jones AR, et al. (2008). The proteome of Toxoplasma gondii: Integration with the genome provides novel insights into gene expression and annotation. Genome Biol 9, R116.

Xia Y, Wang G, Buscariollo D, Pitts RJ, Wenger H, and Zwiebel LJ. (2008). The molecular and cellular basis of olfactory-driven behavior in Anopheles gambiae larvae. Proc Natl Acad Sci USA 105, 6433–6438.

Yates JR, 3rd, Eng JK, and McCormack AL. (1995). Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem 67, 3202–3210.

Zwiebel LJ, and Takken W. (2004). Olfactory regulation of mosquito-host interactions. Insect Biochem Mol Biol 34, 645–652.

Address correspondence to:
Akhilesh Pandey, MD, PhD
McKusick-Nathans Institute of Genetic Medicine
Johns Hopkins University
733 N. Broadway, BRB 527
Baltimore, MD 21205

E-mail: pandey@jhmi.edu

or

Mobolaji A. Okulate, PhD
Department of Natural Sciences
University of Maryland Eastern Shore
Richard Hazel Hall 3063
Princess Anne, Maryland 21833

E-mail: maokulate@umes.edu

Abbreviations Used

CDS = Coding DNA Sequence
GSSP = Genome Search Specific Peptide
MS/MS = Tandem mass spectrometry
PSM = Peptide Spectrum Matches