

An updated catalogue of salivary gland transcripts in the adult female mosquito, *Anopheles gambiae*

Bruno Arcà^{1,2}, Fabrizio Lombardo², Jesus G. Valenzuela³, Ivo M. B. Francischetti³,
Osvaldo Marinotti⁴, Mario Coluzzi² and José M. C. Ribeiro^{3,*}

¹ *Department of Structural and Functional Biology, University “Federico II”, Naples, Italy,* ² *Parasitology Section, Department of Public Health, University “La Sapienza”, Rome, Italy,* ³ *Medical Entomology Section, Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 12735 Twinbrook Parkway, Rockville, MD 20852, USA, and* ⁴ *Department of Molecular Biology and Biochemistry and Department of Biological Chemistry, University of California, Irvine, CA, USA*

Running title: *Anopheles gambiae* sialome

Key words: Saliva, malaria, sialome, transcriptome

* jribeiro@niaid.nih.gov

Abbreviations: AG5, antigen-5 family; EST, expressed sequence tag; nt, nucleotide; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis; UTR, untranslated; H class, housekeeping; NR, non-redundant; RT-PCR, reverse transcriptase-polymerase chain reaction; S class, secreted; T class, transposable element; W class, putative bacterial horizontal transfer; U class, coding for proteins of unknown function.

Summary

Salivary glands of blood-sucking arthropods contain a variety of compounds that prevent platelet and clotting functions and modify inflammatory and immunologic reactions in the vertebrate host. In mosquitoes, only the adult female takes blood meals, while both sexes take sugar meals. With the recent description of the *Anopheles gambiae* genome, and with a set of ~3,000 expressed sequence tags from a salivary gland cDNA library from adult female mosquitoes, we attempted a comprehensive description of the salivary transcriptome of this most important vector of malaria transmission. In addition to many transcripts associated with housekeeping functions, we found an active transposable element, a set of *Wolbachia*-like proteins, several transcription factors including Forkhead, Hairy, and doublesex, extracellular matrix components, and 71 genes coding for putative secreted proteins. Fourteen of these 71 proteins had matching Edman degradation sequences obtained from SDS-PAGE experiments. Overall, 33 transcripts are reported for the first time as coding for salivary proteins. The tissue and sex specificity of these protein-coding transcripts were analyzed by RT-PCR and micro array experiments for insight to their possible function. Notably, 2 gene products appeared to be differentially spliced in the adult female salivary glands, whereas 13 contigs matched predicted intronic regions and may include additional alternatively spliced transcripts. Most *An. gambiae* salivary proteins represent novel protein families of unknown function, potentially coding for pharmacologically or microbiologically active substances. Supplemental data to this work can be found at <http://www.ncbi.nlm.nih.gov/projects/omes/>.

Introduction

The salivary glands of blood-sucking arthropods express a varied mixture of anti-hemostatic and immunomodulatory components that help the arthropod to take, or to find, a blood meal (Ribeiro, 1995). In the case of mosquitoes, only the adult female is hematophagic, whereas both male and females take sugar meals. Perhaps for this reason, adult mosquitoes also have salivary glycosidases (Grossman et al., 1997; Marinotti et al., 1990) and anti-microbials (Rossignol and Lueders, 1986) that may prevent bacterial growth in the sugar meal stored in the mosquito crop.

The “classical” process to learn the function of salivary gland products in vector arthropods starts with the discovery of a biological activity in crude homogenates, then isolation of the protein, and, finally, description of the DNA sequence coding for the protein primary structure. Recent advances in transcriptome techniques led to the reversal of these steps in such a way that the primary sequence of many putatively secreted salivary proteins are now known; but for only a minority of these do we yet know the function and even whether they are really secreted (Ribeiro and Francischetti, 2003). In the case of the mosquito *Anopheles gambiae*, the main vector of malaria in Africa, previous transcriptome analysis of nearly 500 expressed sequence tag (EST) and signal sequence trap methods were used to identify genes expressed in the adult female salivary glands (Arca et al., 1999; Francischetti et al., 2002; Lanfrancotti et al., 2002). Accordingly, a combined non-redundant (NR) set of 40 proteins has been proposed to be of a salivary secretory nature in *An. gambiae*; we can assign a function based on experimental evidence for fewer than 10 of these.

The recent elucidation of the genome of *An. gambiae* associated with high-throughput transcriptome analysis facilitates further gene discovery. In this paper, we present the analysis of an additional set of set of 2,396 salivary gland cDNA sequences (total of 3087 compared to previous set of 691 clones) providing for the discovery of 33 new salivary gland proteins. A NR catalogue

including 73 transcripts—of which 71 code for proteins of a putative secretory nature—is presented and discussed. It should be helpful in designing experiments to determine the function for the majority of these transcripts. Toward this end, we analyzed the tissue and sex specificity of 88 transcripts and found that 27 are either exclusively expressed or enriched in the salivary glands.

Materials and methods

Library construction and cDNA sequencing

An. gambiae salivary gland mRNA was isolated from 80 salivary gland pairs from adult females at days 1 and 2 after emergence using the Micro-FastTrack mRNA isolation kit (Invitrogen, San Diego, CA, USA). A cDNA library was constructed and randomly selected cDNA clones sequenced as previously described (Francischetti et al., 2002).

Bioinformatic tools used

EST were trimmed of primer and vector sequences, clusterized, and compared with other databases as described before (Valenzuela et al., 2003). The BLAST tool (Altschul and Gish, 1996) and CAP3 assembler were used (Huang and Madan, 1999), as well as the ClustalW (Thompson et al., 1994) and Treeview software (Page, 1996). O-glycosylation sites on the proteins were predicted with the program NetOGlyc (<http://www.cbs.dtu.dk/services/NetOGlyc/>) (Hansen et al., 1998). We submitted all translated sequences (starting with a Met) to the Signal P server (Nielsen et al., 1997) to detect signal peptides indicative of secretion. For visualization of EST on the *An. gambiae* genome, the EST and cluster sequences were mapped to the *An. gambiae* genome using the Artemis tool (Berriman and Rutherford, 2003) after downloading the GenBank-formatted files for the

An. gambiae chromosomes from Ensembl found at ftp://ftp.ensembl.org/pub/current_mosquito/data/flatfiles/genbank/. Because the files for each chromosome or chromosome arm are partitioned into several different files, a program written in Visual Basic was used to read the GenBank-format files to obtain a single fasta-formatted file for each chromosome or chromosome arm and a uniform rather than relative location for each gene, producing a feature file that could be read by Artemis.

Accordingly, Artemis could read a single flat file and a single set of features for each chromosome or chromosome arm instead of breaking up each chromosome into several dozen pieces. The unique fasta files for each chromosome were, in turn, broken into 30-kbase fragments with 5 kbase from previous sequence to speed BLAST analysis. The EST and contigs were compared with this fragmented-sequence genomic database by blastn (Altschul et al., 1997) and the output transformed to a file compatible to Artemis using a program written in Visual Basic. Sequence annotation was done with the help of AnoXcel (Ribeiro et al., 2004).

Gel electrophoresis and Edman degradation studies

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) of 20 pairs of homogenized *An. gambiae* adult females salivary glands was performed using 1-mm thick NU-PAGE 4% to 12% gels (Invitrogen). Gels were run with MES buffer according to the manufacturer's instructions and the proteins transferred to a PVDF membrane. The membrane was then stained with Coomassie blue in the absence of acetic acid; visualized bands (including a negative-stained band) were cut and subjected to Edman degradation using a Procise sequencer (Perkin-Elmer Corp., Foster City, CA, USA). More details can be obtained in a previous publication (Francischetti et al., 2002). To find the cDNA sequences corresponding to the amino acid sequence—obtained by Edman degradation of the proteins transferred to PVDF membranes from PAGE gels—we wrote a search program (in Visual Basic) that checked these amino acid sequences

against all possible reading frames of each cDNA sequence obtained in the mass sequencing project. For details, see (Valenzuela et al., 2002b).

Reverse transcription-polymerase chain reaction (RT-PCR) expression analysis

Salivary glands were dissected from 1- to 4-day-old adult females, frozen in liquid nitrogen, and stored at -80°C . Total RNA was extracted from dissected glands, carcasses (i.e. adult females from which salivary glands had been dissected), and adult males using the TRIZOL reagent (Invitrogen) and treated with RNase-free DNase I. DNase-treated total RNA (50 ng) was amplified using the SuperScript one-step RT-PCR system (Invitrogen) according to the manufacturer's instructions. Typically, reverse transcription (50°C , 30 min) and heat inactivation of the reverse transcriptase (94°C , 2 min) were followed by 35 PCR cycles: 30 sec at 94°C , 30 sec at 65°C , 1 min at 72°C . For a subset of primer pairs, the annealing temperature was lowered to $55\text{--}60^{\circ}\text{C}$ for optimal amplification. Twenty-five cycles were used for the amplification of the product based on the ribosomal protein S7 mRNA (*rpS7*) to keep the reaction below saturation levels and to allow a more reliable normalization. The sequences of the oligonucleotide primers used for *rpS7* amplification were: Ag_rpS7-F-5' GGC GAT CAT CAT CTA CGT GC 3' and Ag_rpS7-R-5' GTA GCT GCT GCA AAC TTC GG 3'. The sequences of the other oligonucleotide primers are provided in the supplemental material. Amplification reactions were analyzed on 1.2% agarose gels stained with ethidium bromide.

Microarray analysis

Total RNA of five adult mosquitoes was extracted to prepare each sample. A total of six samples—three of non-blood-fed females, and three from sugar-fed male mosquitoes—were analyzed. Isolated total RNA were processed as recommended by Affymetrix, Inc. (*Affymetrix*

GeneChip Expression Analysis Technical Manual; Affymetrix, Inc., Santa Clara, CA, USA). Data analysis was done with the Gene Chip Operational Software (GCOS) package. Other procedures are exactly as described before (Marinotti et al., 2005). The microarray data are available at <http://www.angagepuci.bio.uci.edu/>.

Results and Discussion

General description of the salivary transcriptome database

Clones (total 3087) were included in the EST salivary database, including 691 previously described (Francischetti et al., 2002). Of these, 176 EST were identified as being of mitochondrial origin according to their match to the *An. gambiae* mitochondrial genome and were not further analyzed. The combined nuclear sequences were assembled into 861 contigs and singletons (in this paper uniformly named contigs) after clusterization of the database (see Materials and methods). To attempt a functional classification of these unique sequences, we compared them with proteome databases by blastx and with protein motifs by rpsblast (see Materials and methods). Following manual annotation of these contigs, which included the assignment of known or putative functions to the translation products, they were further divided into five categories ([Table 1](#)): housekeeping (H class) with 357 contigs and 554 sequences; secreted (S class) with 155 contigs and 1940 sequences; transposable element (T class) with 3 contigs and 3 sequences; putative bacterial horizontal transfer contigs (W class) with 9 contigs and 32 sequences; and a last category composed of contigs coding for proteins of unknown function (U class) with 337 contigs and 382 sequences. Although the S class corresponds to only 18% of the contigs, it consists of 67% of all EST, reflecting

the relatively low complexity and abundance of the secretory material of the organ, as indicated before (Francischetti et al., 2002).

Because a significant number of contigs did not match any protein of the *An. gambiae* proteome set, we considered the possibility that these U class transcripts could be representing mostly untranslated (UTR) mRNA regions. The PCR-based cDNA library used in this work supposedly provides for full-length clones by use of a strategy using polydT primers and a modified polymerase (Zhu et al., 2001) when synthesizing the cDNA from RNA. The cDNA were directionally cloned into the viral vector and sequenced only in the 5'-3' direction, because extensions from the 3' end most often fail to cross the polyA region. Accordingly, clones coding for 5' UTR could derive from full-length clones with unusually long 5' UTR, because our average read is larger than 400 nucleotides (nt). Alternatively, the sequenced cDNA could correspond to the 3' UTR of the transcripts if the polymerase fell off its template in the cDNA synthesis step or in the case of the transcript having a Sfi restriction site, which is used during library construction. When each contig position was located in the *An. gambiae* genome and the closest gene in the same orientation identified, we observed that 230 contigs were near the 3' end of predicted exons, while only 51 were near the 5' region of predicted exons. A χ^2 test indicates this difference to be highly significant ($P < 0.001$). These 230 contigs containing 263 sequences indicated that approximately 9% of the database sequences were truncated. Additionally, 567 contigs overlapped with predicted exon locations, including some that did not give significant blastx matches to the *An. gambiae* proteome because they contained only a few base pairs within the predicted exon. A few (13) contigs ([Supplemental spreadsheet, worksheet "Nuclear", column AR](#)) matched predicted intronic regions, usually on large genes, possibly representing alternative splicing and/or the cloning of unprocessed pre-mRNA. We also observed that many contigs coded for different locations of the same gene. A

non-redundant list of gene matches is provided in the supplemental spreadsheets within the worksheet named "ENSANGP list".

Following visualization of these U-class contigs into the genome using the Artemis tool, we further excluded some potential hits to UTR either because the nearest match was too far from the gene ("too far" was considered a distance longer than the length between start of first exon to end of last exon) or because the contig was probably coding for a novel gene ([25 occurrences; supplemental spreadsheet, worksheet "Nuclear", column AR – search for "novel"](#)). We thus arrived at 177 contigs probably located in the 3' UTR region of predicted genes and 23 located at the 5' UTR. We have also observed that contigs matching 3' UTR tended to be on large genes. Indeed, the set of predicted *An. gambiae* genes identified by direct contig matches to an exon had an average gene length (measured from the beginning of first exon to the end of last exon) of 2733 nt and had 3.32 exons, while that of genes identified by their 3' UTR was more than twice as large (6,112 nt) and with an average of 5.07 exons. Both differences are highly significant ($P < 0.001$) when compared by the Mann-Whitney rank sum test. It should also be considered that some of these putative UTR transcripts may code for not-yet-identified exons. Re-annotation of the database taking into consideration the 3' UTR matches increased significantly the number of probably H class genes while decreasing those of the U class ([Table 1](#)).

Transcribed transposable elements (TE)

Two transcripts on our database (contigs 284 and 285; [Supplemental spreadsheet](#)) possibly derive from transposable elements. Their translation products are similar to those of [Caenorhabditis elegans proteins annotated as CCHC-type and RNA-directed DNA polymerase and integrase](#) and to [TY elements in the GO database](#), and also possess [rve Pfam](#) domains indicative of reverse transcriptases. These transcripts may indicate active ongoing transposition activity in *An. gambiae*.

Transcribed bacteria-like gene products

A relatively large number of transcripts (34 sequences, organized into 11 contigs, representing 1.2 % of the salivary EST originating from nuclear genes) match three genes located contiguously in chromosome arm 3R that code for the putative proteins [ENSANGP00000027299](#), [ENSANGP00000027791](#), and [ENSANGP00000029569](#) (Fig. 1). Investigation of nearby genes identified another instance of a possible family member (Fig. 2; [ENSANGP00000026834](#)) without EST representation in our database. When the program PSI-BLAST was used with protein sequence 29569 above for two iterations, it gave a 0.0-e value with hypothetical protein WD0513 of *Wolbachia* endosymbiont of *Drosophila melanogaster* ([gi:42520378](#)) in addition to identifying all other proteins of the cluster. Further iterations of the program retrieves many bacterial proteins annotated as belonging to the Rhs family (Hill et al., 1994). Although the automatic ENSEMBL translation predictions indicate spliced products for the transcripts coding proteins 27791 and 26834, the cDNA we sequenced did not confirm the predictions. The transcripts are not spliced and show one single large open-reading frame (Fig. 1). The likely single-exon structure of these contiguous genes and their similarity to bacterial proteins suggests that this protein family cluster arose by horizontal transfer from a bacterial genome. Because horizontal gene transfer could be mediated by transposable elements (Syvanen, 1994), we investigated whether such sequences were present in the vicinity of these genes. Indeed, two retro transposable element-like fragments, named TE5P and TE3P, flank the region containing the bacterial genes, together with five additional genes, as shown in Figure 2. TE5P, in particular, is located very close to the 5'-most gene coding for [ENSANGP00000026834](#) and could have originated the lateral transfer. The BLAST alignments of TE5P and TE3P with described transposons are shown in [Fig. 3A](#) and [3B](#). The *Wolbachia* genus consists of rickettsia-like organisms infecting arthropods and conferring the phenomenon of

cytoplasmic incompatibility (Drancourt and Raoult, 1994; Sinkins, 2004). Of interest, Anopheles mosquitoes are resistant to Wolbachia (Kittayapong et al., 2000; Ricci et al., 2002), and it is hypothesized here that these Wolbachia-like transcripts may underlie such resistance.

H class gene products

Putative H class genes were further classified according to their possible function ([Table 2](#)). Results are available online and can be searched on the columns labeled “Class” and “Comments” ([Supplemental spreadsheet](#)). Not surprisingly, the most abundant gene class expressed constitutes members of the protein synthesis machinery, which together with transcription machinery, protein modification, and protein export comprise 34% and 43% of H class contigs and sequences, respectively. Transporters and signal transduction gene products are also highly represented in the library. EST matching transporter proteins were found for several V-type ATPase subunits, Na⁺ + K⁺ ATPases, Ca⁺⁺ ATPases, and several families of solute carriers. V-type ATPases have been implicated in the secretion of saliva in Diptera (Zimmermann et al., 2003). Several transcripts coding putative receptors were also found, including G-coupled proteins ([ENSANGP00000023076](#)), a kinase associated with β-adrenergic receptors ([ENSANGP00000008658](#)), and several subunits of the NMDA/glutamate receptor family ([ENSANGP00000018675](#), [ENSANGP00000025350](#), [ENSANGP00000021195](#)). These may function in the secretion signaling of the salivary glands.

Among the transcripts coding for extracellular matrix components, we highlight those coding for laminin ([ENSANGP00000010745](#), a gene that needs to be corrected in its intron-exon borders), the heparin sulfate proteoglycan perlecan ([ENSANGP00000022422](#)), and the enzyme chondroitin N-acetylgalactosaminyltransferase, involved in the synthesis of extracellular mucopolysaccharides ([ENSANGP00000020105](#)). These extracellular constituents may be important for Plasmodium

recognition of the salivary glands, because sporozoites are known to recognize sulfated polysaccharides (Pinzon-Ortiz et al., 2001).

Several transcripts matched genes coding for transcription factors. [Table 3](#) lists some of interest for the specialized function of the female salivary gland, including transcription factors associated with expression of ER chaperones (XBP-1), general transcription factors, and those associated with tissue differentiation. In particular, two genes coding for Forkhead transcription factors are indicated, as well as three involved in the Hairy pathway. The Forkhead and Hairy transcription factors have been implicated in *Drosophila* salivary gland differentiation and salivary protein expression (Lee and Frasch, 2004; Mach et al., 1996; Myat and Andrew, 2000; Myat and Andrew, 2002; Myat et al., 2000; Poortinga et al., 1998). Expression of the gene coding for doublesex, which is associated with sex-specific gene expression in *Drosophila* (Baker et al., 1989; Baker and Wolfner, 1988) is a good candidate to explain the sexual dimorphism observed in adult mosquito salivary glands.

Updated catalogue of putative secreted salivary proteins

After identifying putative secreted proteins ([Supplemental spreadsheet](#)), we used this data set and the Artemis tool to identify novel proteins coded in the *An. gambiae* genome. Indications of secreted polypeptides were obtained with searches for the presence of signal peptides with the SignalP program (Nielsen et al., 1997) and of O-linked galactosylation sites (indicative of mucins) with the NetOGlyc program (Hansen et al., 1998). A NR set of 73 putative proteins expressed in the salivary glands of *An. gambiae* is presented in [Table 4](#); it includes 71 polypeptides predicted as secretory, 40 of which have been described previously. Of these 40, 7 are described now in full length. Thirty-three proteins are indicated for the first time to be expressed in the salivary glands of adult female mosquitoes. Of these 33 proteins, 29 were predicted by the ENSEMBL annotation

pipeline and 4 are novel. Of the 29 predicted by ENSEMBL, 17 were re-annotated to fix the starting Met or stop codons.

D7 salivary proteins: The first member of the D7 protein family was described in the mosquito *Aedes aegypti* (James et al., 1991) and later found in virtually all mosquito sialotranscriptomes. Short (~15 kDa) and long (~30 kDa) forms are recognized. Long D7 forms also exist in sand flies (Valenzuela et al., 2002a). The function of these proteins has not been verified, although one short D7 protein from *An. stephensi*, named hamadarin, was shown to prevent kallikrein activation by Factor XIIa (Isawa et al., 2002). Previously, one long D7 and five short D7 proteins were known in *An. gambiae* (Arca et al., 2002; Francischetti et al., 2002). [Table 4](#) shows these six proteins and two additional D7 proteins, all coded from contiguous genes in chromosome arm 3R. The three long D7 genes follow each other in the forward direction, the first two having four exons, but the last having only three exons ([Fig. 4](#)). The five short-form genes follow the long D7 cassette in reverse orientation, the first four having three exons, while the fifth has only two exons. Notably, an apparently non-coding transcript (contig_709) maps just 250 nt downstream of the short D7 cassette in the reverse orientation of the gene. We speculate that this transcript may be associated with regulation of D7 expression. Combined, these genes represented 574 sequences in our database, or nearly 20% of over 3,000 EST. It is also interesting to note that the last gene in each of the cassettes, i.e. D7L3 and D7r5, were the least represented in number of EST, indicating that they are expressed at lower levels than their similar neighbouring genes. Moreover, in comparison to the other members of the cluster, D7L3 and D7r5 differ in the number of exons and the pattern of expression not restricted to female salivary glands ([Table 4](#)). Evidence for the synthesis of all but one of these proteins (named D7L3 in [Table 4](#)) in the salivary glands of female mosquitoes was found by Edman degradation of bands resolved by SDS-PAGE.

Antigen 5 (AG5) family: Four genes coding for members of the AG5 family were identified by matching salivary transcripts. AG5-related salivary products are members of a group of secreted proteins that belong to the CAP family (cysteine-rich secretory proteins; AG5 proteins of insects; pathogenesis-related protein 1 of plants) (Megraw et al., 1998). The CAP family is related to venom allergens in social wasps and ants (Hoffman, 1993; King and Spangfort, 2000) and to antifungal proteins in plants (Stintzi et al., 1993; Szyperski et al., 1998). Members of this protein family are found in the salivary glands of many blood-sucking insects (Francischetti et al., 2002; Li et al., 2001; Valenzuela et al., 2002b). These animal proteins have no known function, except for a few cases: one *Conus* protein was recently shown to have proteolytic activity (Milne et al., 2003), snake venom proteins of this same family have been shown to contain smooth muscle-relaxing activity (Yamazaki et al., 2002; Yamazaki and Morita, 2004), and the salivary neurotoxin of the venomous lizard *Heloderma horridum* is also a member of this protein family (Nobile et al., 1996). Three of the four genes identified by the transcriptome are located in a cluster of genes in chromosome arm 2L, with these three genes receiving 106, 1, and 1 matches from EST (Table 4). The fourth gene, located on chromosome arm 2R, received only one EST match from our database. Of these four putative protein sequences, one is novel and another reports the full length of a previously identified salivary-expressed sequence. Further work on this gene family will be reported elsewhere (B. Arcà et al., manuscript in preparation).

The SG1 family of Anopheline proteins: This family of salivary proteins, having mature molecular weight near 44 kDa, was previously described as SG1 or gSG1 (Arca et al., 1999; Lanfrancotti et al., 2002). They do not yield significant similarities by blastp to other proteins in the NCBI database except for other anopheline proteins, including the distantly related TRIO protein. Six genes of this family are known in *An. gambiae*, five of which reside in chromosome X, while the gene coding for the TRIO protein is in the 2R chromosome arm ([Supplemental Table 4](#)). Two of

these gene products are reported here in their full-length configuration. Four of the five genes in the X chromosome are observed in a tandem configuration, including one in reverse orientation ([Fig. 5](#)). This family has a relatively high EST representation in our database, with a total of 114 EST. Except for gSG1a having two EST, all others had ten or more EST represented (Table 4). One polyadenylated transcript (contig_78) was found mapping in anti-sense orientation of SG1_like-3 ([Fig. 5](#)). Its possible significance is unknown. Alignment of the six protein sequences is not very informative, except for a weak similarity region in the middle of the protein ([Fig. 6](#)). A hidden Markov model made from the Clustal alignments of the six proteins was used to search the NR protein database of NCBI. All retrieved protein sequences were of anopheline origin (not shown). Evidence for secretion of gSG1b, SG1, and SG1-like3_long was found by Edman degradation of SDS-PAGE protein bands ([Supplemental Table 4](#)).

Mucins: Due to their putative high number of serine and threonine and their high probability of having ten or more O-linked N-acetylgalactosamine, three proteins are identified as mucins, two of which have been described previously and one of which is novel ([Table 4](#)). All of these proteins have homologues found in sialotranscriptomes of *An. stephensi*, and one in *Culex quinquefasciatus*. These proteins might function in the lubrication of the mosquito mouthparts. One of them (SG3) has a weak indication of a chitin binding site domain, and it is possible that it binds to the chitinous linings of the salivary ducts and mouthparts. These predicted transcripts were found to be enriched in salivary glands of females and were also found in male mosquitoes ([Table 4](#)).

Other salivary-expressed genes coding for proteins or peptides of unknown function: Table 4 lists 33 peptides and proteins with no hits or non-significant e values when compared with GO, PFAM, and SMART databases. Fifteen of them were not previously reported as being expressed in the salivary glands of *An. gambiae*. Additionally, three previously described messages are now reported in their full CDS form. Thirty of these 33 proteins have a signal peptide indicative of

secretion, although it should be noted that their final destination could be the ER or Golgi complex. Except for the transcript coding for the protein described before as cE5 (Arca et al., 1999), which is the homologue of the *An. albimanus* antithrombin peptide named anophelin (Valenzuela et al., 1999), we have no information that could indicate the function of these gene products. Some of these proteins apparently result from gene duplication events, such as those listed in Table 4 as: 1) hyp15 and hyp17, coding for basic peptides (pI > 11.0) of ~4.7 kDa and residing contiguously on chromosome X; 2) hyp10 and hyp12, coding for slightly acidic peptides of ~7.5 kDa residing contiguously on chromosome arm 3R; 3) hyp8.2 and hyp6.2, apparently unrelated in sequence similarity, but coding for mature peptides of 7.6 and 6.2 kDa and residing contiguously on chromosome arm 2L; 4) SG2 and SG2A proteins, coding for mature peptides apparently unrelated in sequence similarity of 9.5 and 15.5 kDa, residing close to each other in 2L; and 5) hyp4.2 and hyp13, coding for mature peptides of 4.2 and 3.6 kDa on chromosome arm 2R. These pairs of genes show identical or very similar patterns of expression (see [Table 4](#)) and possibly reflect examples of gene duplication and divergence of function (Sankoff, 2001), as do the D7, SG1, and AG5 families described above.

Of these 33 salivary gland-expressed genes, 22 appear to code for proteins found only in anopheline mosquitoes, 5 are common to Culicidae, 1 is known to occur also in *Drosophila*, and 4 are more generally conserved. Together with the 6 members of the SG1 family, there are altogether 28 gene products that appear to be unique to anophelines and could be used as antigenic markers of anopheline exposure for epidemiologic studies as done before for ticks and sand flies (Barral et al., 2000; Schwartz et al., 1990).

Among the gene products unique to mosquitoes (including anophelines and culicines), we report here the full-length information for the 30_kD protein. The 30_kD transcript sequence produced nearly 200 EST matches from our database, being the second most abundantly expressed

gene in the salivary glands of *An. gambiae*. Splice variants of this protein are [apparent](#) from the different assemblies of these EST. Transcripts coding for members of this acidic protein family, first identified as the 30-kDa Aedes allergen (Simons and Peng, 2001) and also named GE-rich protein (Valenzuela et al., 2003), were found in all previously described transcriptomes of both culicine and anopheline mosquitoes. Another uniquely mosquito protein family is represented by hyp55.3 ([Table 4](#)). Additionally, two *An. gambiae* genes code for a protein similar to a salivary Culex protein annotated as putative 14.5-kDa salivary peptide. Although these two Anopheles proteins appear to be related, their corresponding genes are located in different chromosomes. The protein indicated as SG2a also has homology to a Culex putative salivary protein.

One single EST identified an *An. gambiae* gene coding for a protein with 49% identity to Drosophila retinin, a protein of unknown function expressed in the insect eye. Genes of a more general conserved nature expressed in *An. gambiae* salivary glands include the previously described selenoprotein, the hypothetical proteins named in Table 4 as hyp14.6 and hyp1.2, and calreticulin. Although calreticulin functions as a chaperone in the ER and the *An. gambiae* salivary calreticulin has a carboxyterminal sequence HDEL suggestive of retention in the ER, proteins of this family have anti-thrombotic functions in the extracellular compartment (Nash et al., 1994; Nauseef et al., 1995; Pike et al., 1998; Sontheimer et al., 1995) and have been described in the saliva of ticks (Jaworski et al., 1995). Its possible function in the saliva of ticks and Anopheles mosquitoes remains to be investigated.

For this broad class of proteins, we found evidence of synthesis of gSG7, gSG7-2 and the 30_kDa peptides by Edman degradation of salivary gland peptides resolved by SDS-PAGE.

Enzymes: Several transcripts coding for enzymes are identifiable. These enzymes are probably associated with four groups of functional activities: 1) Catabolism of hemostasis and inflammation mediators, including the previously described 5'-nucleotidase and apyrase, that may

facilitate acquisition of blood meals by removing pharmacologically active nucleotides at the site of the injury (Ribeiro and Francischetti, 2003). We found strong Edman degradation signal in SDS-PAGE bands matching the predicted amino terminal of the putative 5'-nucleotidase. Additionally, we describe a salivary peroxidase homologous to the enzyme in *An. albimanus* that acts as a vasodilator by its catechol-oxidase activity (Ribeiro and Valenzuela, 1999; Ribeiro and Nussenzveig, 1993). Transcripts for enzymes of this class were well expressed, with 74, 26, and 6 EST found in the database for the 5'-nucleotidase, apyrase, and peroxidase, respectively. We also report a gene product coding for an epoxy hydrolase, represented by a single EST match. The corresponding gene, as presently annotated, codes for a truncated protein ([ENSANGP00000008689](#)) for which we cannot find the starting methionine and thus infer whether it may be secreted. Notably, this gene is located contiguously and in reverse orientation to the salivary apyrase gene in chromosome arm 3L. It is included in the detailed analysis because of the potential role of arachidonic acid epoxides and epoxide hydrolase in inflammation and hemostasis (Spector et al., 2004; Weintraub et al., 1999). 2) Sugar digestion-related enzymes, including amylase and α -glucosidase, both described in their full-length CDS in this paper. 3) Proteases, including four serine proteases that could be involved with specific host proteolytic events that could affect clotting or the complement cascade. Two other serine proteases are probably involved in immunity, as they are similar to prophenoloxidase-activating enzymes. One metalloprotease homologous with a *Drosophila* enzyme involved in remodeling of the salivary glands was also found in the salivary transcriptome of *An. gambiae*. We could not ascertain its full-length coding sequence and therefore the occurrence of a signal peptide, but it is included here for having seven putative O-glycosylation sites and as a reference to future studies regarding its involvement in the development of the adult salivary gland or as a potentially secreted metalloprotease, as occurs with ticks (Francischetti et al., 2003). 4) The homologue of *Drosophila* peroxinectin is also identified. Peroxinectins are

multifunctional proteins with a peroxidase domain and cell adhesion activity. This enzyme may have a role in blood feeding or, more probably, may be involved with sclerotization of extracellular matrix constituents. We found strong Edman degradation signal in SDS-PAGE bands matching the predicted amino terminal of the putative 5'-nucleotidase.

Immunity-related gene products: Transcripts coding for two different lysozymes were found, one abundantly transcribed (36 transcripts) and one with only one EST. Lysozyme activity was previously shown to be abundant in salivary glands of both male and female mosquitoes, where it may help contain microbial growth in stored sugar meals (Moreira-Ferro et al., 1998; Rossignol and Lueders, 1986). Two lectin-coding genes were also identified as expressed in the salivary glands: *i*) a C-type lectin, found through a match to the 3'-UTR of a mRNA coding a putative protein with PFAM and SMART domains, indicating this type of lectin, corresponding to [ENSANGP00000020547](#); and *ii*) galectin, found by one EST match, corresponding to [ENSANGP00000026948](#). Although the encoded *An. gambiae* galectin does not have a signal peptide indicative of secretion, it is known that galectins may be secreted via an alternative mechanism (Nickel, 2003). If secreted, these proteins may be responsible for the hemagglutinating activity of *Anopheles saliva* (Metcalf, 1945), which can help in concentrating the blood meal (Vaughan et al., 1991).

Tissue and sex transcriptions specificity

While the function of most putative salivary proteins and peptides is presently unknown, determination of their tissue and sex specificity may help to direct further research to characterize these gene products. To this end, we used RT-PCR on total RNA extracted from female salivary glands, female carcasses (i.e. adult females from which salivary glands had been removed), and adult males. Eighty-eight mRNA, mostly encoding secreted polypeptides, were selected on the basis of

either their sequence similarity or absence of any similarity to known proteins. Nine additional mRNA previously analysed by RT-PCR (AgApy, AgApyL1, D7r1, gSG1, gSG3, gSG6, gSG7, gSG10, cE5/anophelin) were also included as controls for the amplification reactions (Arca et al., 1999; Francischetti et al., 2002; Lanfrancotti et al., 2002). We have also analysed the sex-dependent expression of the genes shown in [Table 4](#) using the Affymetrix microarray chip for comparison with the RT-PCR results. Of the 72 gene products shown in Table 4, 3 are not represented in the Affymetrix gene set. The combination of these two independent analyses allows the delineation of three categories. The first is represented by genes either expressed at approximately the same level in the three tissues examined or less abundantly transcribed in female glands: proteins encoded by these ubiquitous genes are presumably involved in housekeeping functions. Approximately one third of the genes analysed (25/72) belong to this group; a few representatives are shown in [Fig. 7A](#). In the microarray experiment, these genes should show equal expression in males and females, and indeed, the average of the log of the hybridization signal ratio between sugar-fed females and males for this set of genes was 0.11 ± 0.07 (mean \pm SE; $n = 24$), a result not significantly different from 0 (a value indicating equal hybridization signal, because $10^0 = 1$). These ubiquitously expressed genes may play housekeeping roles related to glandular function, such as calreticulin, as well as general immune mechanisms, such as the case of the lysozymes and prophenoloxidase activating serine proteases.

The second category of tissue-specific expression consists of 34 genes represented by 17 transcripts that are either female salivary gland specific (marked as "SG" in Table 4) ([Fig. 5B](#)) or whose expression is enriched in the female salivary glands (the 17 additional genes are marked as "[Enrich](#)" in Table 4) ([Fig. 7C](#)): Microarray experiments indicated a highly significant differential ratio of expression between sexes in both subgroups. The SG set had a mean log of ratios of 1.45 ± 0.22 ($n = 17$), and the Enrich set had an average ratio of 1.24 ± 0.19 ($n = 15$; two genes are missing on the microarray chip), indicating a geometric average increase of transcript expression of

28 and 17 fold, respectively, when female transcript expression is compared with that of males. Additionally, the hybridization signals for each probe set were analyzed with an algorithm (GCOS) that designated the presence or absence of the corresponding transcript. Again, in most cases, the transcripts identified as female specific by RT-PCR were confirmed by the microarray data.

The female-enriched or female-specific salivary gland genes are likely to play a role in blood feeding as anti-hemostatics or immunomodulators. Nineteen of the 34 genes are either newly described here or had not been previously analyzed for their expression specificity. Jointly, these genes include all eight members of the D7 family, the AG5 protein gvag, all six members of the G1 family, and the protein/peptides named 30_kDa, gSG7-2, hyp17, gSG7, hyp8.2, gsg6, hyp14.5, gSG5, hyp15, gsg8, hyp37.7, hyp6.2, the enzyme salivary peroxidase, 5/-nucleotidase, apyrase, the serine proteases sal_serpro1, sal_tryp_XII, Sal_serpro2, the salivary epoxy hydrolase, and the salivary galectin.

The third subgroup of genes includes those expressed in female salivary glands as well as in adult males, with absent or irrelevant expression in female carcasses ([Fig. 7D](#)). Microarray experiments indicated that the average log of the ratio of the hybridization intensities between females and males was not significantly different from 0 (-0.13 ± 0.08 ; $n = 12$). We assume that these genes are salivary gland specific and expressed both in male and female glands. The corresponding gene products are likely involved in sugar feeding, antimicrobial activity, or in more general physiologic gland functions. Overall, 13 genes appear to be part of this group, including those encoding 3 mucins, and the proteins/peptides hyp55.3, hy10, hyp12, sg2, sg2a, gSG9, hyp6.3, and the sugar-digesting enzymes amylase and maltase ([Table 4](#)). With the exception of enzymes that may help sugar digestion in the mosquito crop and midgut and the mucins that might help maintenance of the food canal, the function of the remaining gene products of this group is presently unknown.

Although the results obtained from the microarray experiment on average agree well with those from the RT-PCR experiments, we found some noticeable discrepancies in the results for the genes *sal_galectin*, *hyp14.5*, *Sal_serpro2*, *sal_tryp_XII*, *Ag_epoxy_hydrolase*, and *ag5-related-4*. One of the reasons for the observed incongruity may be related to alternative splicing of the gene products. Indeed, RT-PCR expression analysis followed by cloning and sequencing of the amplified fragments suggested that the salivary galectin (not shown) and *sal_tryp_XII* genes produce different polypeptides by alternative splicing. In the case of female gland-specific *sal_tryp_XII*, a band 126 bp in length is obtained, as expected, when female salivary gland RNA is used as template, whereas larger products are amplified from RNA extracted from carcasses and males (Fig. 8). Sequence analysis indicated the longer product to be a transcript that retains a 98-bp intron carrying an in-frame stop codon. This would give rise to a hypothetical truncated product of 112 amino acids in place of the putative trypsin-like protease produced in female salivary glands (431 amino acids). It is possible that this tissue- and sex-specific splicing may have a regulatory role, producing a functional protease only in the saliva of *An. gambiae* females. As suggested above, if secreted, it may influence the clotting and/or the complement cascades of vertebrate hosts. Microarray hybridization experiments using a set of 10 or 11 short probes as used in the Affymetrix chip cannot distinguish between these splice variants, making such comparisons invalid (Carter et al., 2005). On the other hand, the incongruence between RT-PCR experiments and microarray data may point to differentially spliced genes that may have importance in tissue translation selectivity (Black, 2003).

Two of the nine genes included in this analysis as controls showed an expression pattern slightly different from what has been reported before, most probably because new primer pairs and different amplification conditions were employed. The expression of *gSG7* appeared enriched in female glands, rather than equally expressed in female salivary glands and in males as previously reported (Lanfrancotti et al., 2002), whereas *cE5* showed very little expression in carcasses in

comparison to what was observed before (Arca et al., 1999). In Table 4, they have been classified according to the more recent RT-PCR results, although it should be kept in mind that their expression pattern may be in between these categories. Very good overlap with previous analyses was obtained with the other seven genes used as controls (AgApy, AgApyL1, D7r1, gSG1, gSG3, gSG6, and gSG10). It is also interesting that cE5 is a homologue of anophelin, a potent anti-thrombin peptide found in the salivary glands of the New World mosquito *An. albimanus* (Francischetti et al., 1999; Valenzuela et al., 1999); however, in *An. gambiae*, cE5 is not found selectively in female glands, as shown here and previously (Arca et al., 1999), raising the possibility that it may exert a different function in Old World mosquitoes.

Concluding remarks

Using high-throughput transcriptome analysis, we significantly expanded the *An. gambiae* salivary gland transcript repertoire. Thirty-three novel putative salivary proteins were identified, and the full-length sequences of seven previously identified partial cDNA were reported. Moreover, tissue-specific expression studies on selected clones allowed us to identify 27 additional genes that are either enriched or specifically expressed in the salivary glands. The information obtained in the course of this analysis, joined to the results from previous studies, allowed us to compile an updated catalogue that includes a total of 72 transcripts mainly encoding putative secreted products. Forty-seven of these transcripts encode proteins that may play essential physiologic roles, as indicated by their exclusive or preferential expression in female and/or male salivary glands. This catalogue makes the mosquito *An. gambiae* the arthropod disease vector for which the most complete salivary transcriptome is available. On the other hand, the fraction of genes included in this list for which we know or can postulate a function is surprisingly small, emphasizing how much we still have to learn about bioactive molecules from the saliva of blood-feeding arthropods. We believe that

this updated catalogue should help our continuing effort of understanding the evolution of blood sucking in vector arthropods and the discovery of novel pharmacologically active compounds.

We are grateful to Brenda Marshal for editorial assistance. This work was supported in part by grants from the European Union to M.C. and B.A. (BioMalPar N 503578), MIUR/COFIN funds to Vincenzo Petrarca and B.A., the Intramural Research Program of the National Institute of Allergy and Infectious Diseases, National Institutes of Health to JMCR, and by the UND/World Bank/WHO Special Programme for Research and Training in Tropical Diseases (TDR), ID A20314 to OM.

References

- Altschul, S. F. and Gish, W.** (1996). Local alignment statistics. *Methods Enzymol* **266**, 460-80.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J.** (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389-402.
- Arca, B., Lombardo, F., de Lara Capurro, M., della Torre, A., Dimopoulos, G., James, A. A. and Coluzzi, M.** (1999). Trapping cDNAs encoding secreted proteins from the salivary glands of the malaria vector *Anopheles gambiae*. *Proc. Natl. Acad. Sci. U S A* **96**, 1516-21.
- Arca, B., Lombardo, F., Lanfrancotti, A., Spanos, L., Veneri, M., Louis, C. and Coluzzi, M.** (2002). A cluster of four D7-related genes is expressed in the salivary glands of the African malaria vector *Anopheles gambiae*. *Insect Mol Biol* **11**, 47-55.
- Baker, B. S., Burtis, K., Goralski, T., Mattox, W. and Nagoshi, R.** (1989). Molecular genetic aspects of sex determination in *Drosophila melanogaster*. *Genome* **31**, 638-45.
- Baker, B. S. and Wolfner, M. F.** (1988). A molecular analysis of doublesex, a bifunctional gene that controls both male and female sexual differentiation in *Drosophila melanogaster*. *Genes Dev* **2**, 477-89.
- Barral, A., Honda, E., Caldas, A., Costa, J., Vinhas, V., Rowton, E. D., Valenzuela, J. G., Charlab, R., Barral-Netto, M. and Ribeiro, J. M.** (2000). Human immune response to sand fly salivary gland antigens: a useful epidemiological marker? *Am. J. Trop. Med. Hyg.* **62**, 740-5.
- Berriman, M. and Rutherford, K.** (2003). Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4**, 124-32.
- Black, D. L.** (2003). Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**, 291-336.
- Carter, S. L., Eklund, A. C., Mecham, B. H., Kohane, I. S. and Szallasi, Z.** (2005). Redefinition of Affymetrix probe sets by sequence overlap with cDNA microarray probes reduces cross-platform inconsistencies in cancer-associated gene expression measurements. *BMC Bioinformatics* **6**, 107.
- Drancourt, M. and Raoult, D.** (1994). Taxonomic position of the rickettsiae: current knowledge. *FEMS Microbiol Rev* **13**, 13-24.
- Francischetti, I. M., Mather, T. N. and Ribeiro, J. M.** (2003). Cloning of a salivary gland metalloprotease and characterization of gelatinase and fibrin(ogen)lytic activities in the saliva of the Lyme disease tick vector *Ixodes scapularis*. *Biochem Biophys Res Commun* **305**, 869-75.
- Francischetti, I. M., Valenzuela, J. G., Pham, V. M., Garfield, M. K. and Ribeiro, J. M.** (2002). Toward a catalog for the transcripts and proteins (sialome) from the salivary gland of the malaria vector *Anopheles gambiae*. *J. Exp. Biol.* **205**, 2429-51.
- Francischetti, I. M., Valenzuela, J. G. and Ribeiro, J. M.** (1999). Anophelin: kinetics and mechanism of thrombin inhibition. *Biochemistry* **38**, 16678-85.
- Grossman, G. L., Campos, Y., Severson, D. W. and James, A. A.** (1997). Evidence for

two distinct members of the amylase gene family in the yellow fever mosquito, *Aedes aegypti*. *Insect Biochem. Mol. Biol.* **27**, 769-81.

Hansen, J. E., Lund, O., Tolstrup, N., Gooley, A. A., Williams, K. L. and Brunak, S. (1998). NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surface accessibility. *Glycoconj J* **15**, 115-30.

Hill, C. W., Sandt, C. H. and Vlazny, D. A. (1994). Rhs elements of *Escherichia coli*: a family of genetic composites each encoding a large mosaic protein. *Mol Microbiol* **12**, 865-71.

Hoffman, D. R. (1993). Allergens in Hymenoptera venom. XXV: The amino acid sequences of antigen 5 molecules and the structural basis of antigenic cross-reactivity. *J Allergy Clin Immunol* **92**, 707-16.

Huang, X. and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res* **9**, 868-77.

Isawa, H., Yuda, M., Orito, Y. and Chinzei, Y. (2002). A mosquito salivary protein inhibits activation of the plasma contact system by binding to factor XII and high molecular weight kininogen. *J. Biol. Chem.* **13**, 13.

James, A. A., Blackmer, K., Marinotti, O., Ghosn, C. R. and Racioppi, J. V. (1991). Isolation and characterization of the gene expressing the major salivary gland protein of the female mosquito, *Aedes aegypti*. *Mol Biochem Parasitol* **44**, 245-53.

Jaworski, D. C. S. F. A., Lamoreaux, W., Coons, L. B., Muller, M. T. and Needham, G. R. (1995). A secreted calreticulin protein in ixodid tick (*Amblyomma americanum*) saliva. *J. Insect Physiol.* **41**, 369-375.

King, T. P. and Spangfort, M. D. (2000). Structure and biology of stinging insect venom allergens. *Int Arch Allergy Immunol* **123**, 99-106.

Kittayapong, P., Baisley, K. J., Baimai, V. and O'Neill, S. L. (2000). Distribution and diversity of *Wolbachia* infections in Southeast Asian mosquitoes (Diptera: Culicidae). *J Med Entomol* **37**, 340-5.

Lanfrancotti, A., Lombardo, F., Santolamazza, F., Veneri, M., Castrignano, T., Coluzzi, M. and Arca, B. (2002). Novel cDNAs encoding salivary proteins from the malaria vector *Anopheles gambiae*. *FEBS Lett.* **517**, 67-71.

Lee, H. H. and Frasch, M. (2004). Survey of forkhead domain encoding genes in the *Drosophila* genome: Classification and embryonic expression patterns. *Dev Dyn* **229**, 357-66.

Li, S., Kwon, J. and Aksoy, S. (2001). Characterization of genes expressed in the salivary glands of the tsetse fly, *Glossina morsitans morsitans*. *Insect Mol. Biol.* **10**, 69-76.

Mach, V., Ohno, K., Kokubo, H. and Suzuki, Y. (1996). The *Drosophila* fork head factor directly controls larval salivary gland-specific expression of the glue protein gene *Sgs3*. *Nucleic Acids Res* **24**, 2387-94.

Marinotti, O., James, A. and Ribeiro, J. M. C. (1990). Diet and salivation in female *Aedes aegypti* mosquitoes. *J. Insect Physiol.* **36**, 545-548.

Marinotti, O., Nguyen, Q. K., Calvo, E., James, A. A. and Ribeiro, J. M. C. (2005). Microarray analysis of genes showing variable expression following a bloodmeal in *Anopheles gambiae*. *Insect Mol. Biol.* **14**, 365-374.

Megraw, T., Kaufman, T. C. and Kovalick, G. E. (1998). Sequence and expression of *Drosophila* Antigen 5-related 2, a new member of the CAP gene family. *Gene* **222**, 297-304.

Metcalf, R. L. (1945). The physiology of the salivary glands of *Anopheles quadrimaculatus*. *J. Nat. Malaria Soc.* **4**, 271.

- Milne, T. J., Abbenante, G., Tyndall, J. D., Halliday, J. and Lewis, R. J.** (2003). Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily. *J Biol Chem* **278**, 31105-10.
- Moreira-Ferro, C. K., Daffre, S., James, A. A. and Marinotti, O.** (1998). A lysozyme in the salivary glands of the malaria vector *Anopheles darlingi*. *Insect Mol Biol* **7**, 257-64.
- Myat, M. M. and Andrew, D. J.** (2000). Fork head prevents apoptosis and promotes cell shape change during formation of the *Drosophila* salivary glands. *Development* **127**, 4217-26.
- Myat, M. M. and Andrew, D. J.** (2002). Epithelial tube morphology is determined by the polarized growth and delivery of apical membrane. *Cell* **111**, 879-91.
- Myat, M. M., Isaac, D. D. and Andrew, D. J.** (2000). Early genes required for salivary gland fate determination and morphogenesis in *Drosophila melanogaster*. *Adv Dent Res* **14**, 89-98.
- Nash, P. D., Opas, M. and Michalak, M.** (1994). Calreticulin: not just another calcium-binding protein. *Mol. Cell. Biochem.* **135**, 71-78.
- Nauseef, W. M., McCormick, S. J. and Clark, R. A.** (1995). Calreticulin functions as a molecular chaperone in the biosynthesis of myeloperoxidase. *J. Biol. Chem.* **270**, 4741-4747.
- Nickel, W.** (2003). The mystery of nonclassical protein secretion. A current view on cargo proteins and potential export routes. *Eur J Biochem* **270**, 2109-19.
- 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 Eng.* **10**, 1-6.
- Nobile, M., Noceti, F., Prestipino, G. and Possani, L. D.** (1996). Helothermine, a lizard venom toxin, inhibits calcium current in cerebellar granules. *Exp Brain Res* **110**, 15-20.
- Page, R. D.** (1996). TreeView: an application to display phylogenetic trees on personal computers. *Comput. Appl. Biosci.* **12**, 357-8.
- Pike, S. E., Yao, L., Jones, K. D., Cherney, B., Appella, E., Sakaguchi, K., Nakhasi, H., Teruya-Feldstein, J., Wirth, P., Gupta, G. et al.** (1998). Vasostatin, a calreticulin fragment, inhibits angiogenesis and suppresses tumor growth. *J Exp Med* **188**, 2349-56.
- Pinzon-Ortiz, C., Friedman, J., Esko, J. and Sinnis, P.** (2001). The binding of the circumsporozoite protein to cell surface heparan sulfate proteoglycans is required for plasmodium sporozoite attachment to target cells. *J Biol Chem* **276**, 26784-91.
- Poortinga, G., Watanabe, M. and Parkhurst, S. M.** (1998). *Drosophila* CtBP: a Hairy-interacting protein required for embryonic segmentation and hairy-mediated transcriptional repression. *Embo J* **17**, 2067-78.
- Ribeiro, J. M. and Francischetti, I. M.** (2003). Role of arthropod saliva in blood feeding: sialome and post-sialome perspectives. *Annu. Rev. Entomol.* **48**, 73-88.
- Ribeiro, J. M., Topalis, P. and Louis, C.** (2004). Anoxcel: an *Anopheles gambiae* protein database. *Insect Mol Biol* **13**, 449-57.
- Ribeiro, J. M. and Valenzuela, J. G.** (1999). Purification and cloning of the salivary peroxidase/catechol oxidase of the mosquito *Anopheles albimanus*. *J. Exp. Biol.* **202**, 809-16.
- Ribeiro, J. M. C.** (1995). Blood-feeding arthropods: Live syringes or invertebrate pharmacologists? *Infect. Agents Dis.* **4**, 143-152.
- Ribeiro, J. M. C. and Nussenzeig, R. H.** (1993). The salivary catechol oxidase/peroxidase activities of the mosquito, *Anopheles albimanus*. *J. Exp. Biol.* **179**, 273-287.
- Ricci, I., Cancrini, G., Gabrielli, S., D'Amelio, S. and Favi, G.** (2002). Searching for

Wolbachia (Rickettsiales: Rickettsiaceae) in mosquitoes (Diptera: Culicidae): large polymerase chain reaction survey and new identifications. *J Med Entomol* **39**, 562-7.

Rossignol, P. A. and Lueders, A. M. (1986). Bacteriolytic factor in the salivary glands of *Aedes aegypti*. *Comp. Biochem. Physiol.* **83B**, 819-822.

Sankoff, D. (2001). Gene and genome duplication. *Curr. Opin. Genet. Dev.* **11**, 681-4.

Schwartz, B. S., Ribeiro, J. M. and Goldstein, M. D. (1990). Anti-tick antibodies: an epidemiologic tool in Lyme disease research. *Am J Epidemiol* **132**, 58-66.

Simons, F. E. and Peng, Z. (2001). Mosquito allergy: recombinant mosquito salivary antigens for new diagnostic tests. *Int. Arch. Allergy Immunol.* **124**, 403-5.

Sinkins, S. P. (2004). Wolbachia and cytoplasmic incompatibility in mosquitoes. *Insect Biochem Mol Biol* **34**, 723-9.

Sontheimer, R. D., Nguyen, T. Q., Cheng, S. T., Lieu, T. S. and Capra, J. D. (1995). The unveiling of calreticulin - A clinically relevant tour of modern cell biology. *J. Investig. Med.* **43**, 362-370.

Spector, A. A., Fang, X., Snyder, G. D. and Weintraub, N. L. (2004).

Epoxyeicosatrienoic acids (EETs): metabolism and biochemical function. *Prog Lipid Res* **43**, 55-90.

Stintzi, A., Heitz, T., Prasad, V., Wiedemann-Merdinoglu, S., Kauffmann, S., Geoffroy, P., Legrand, M. and Fritig, B. (1993). Plant 'pathogenesis-related' proteins and their role in defense against pathogens. *Biochimie* **75**, 687-706.

Syvanen, M. (1994). Horizontal gene transfer: evidence and possible consequences. *Annu Rev Genet* **28**, 237-61.

Szyperski, T., Fernandez, C., Mumenthaler, C. and Wuthrich, K. (1998). Structure comparison of human glioma pathogenesis-related protein GliPR and the plant pathogenesis-related protein P14a indicates a functional link between the human immune system and a plant defense system. *Proc Natl Acad Sci U S A* **95**, 2262-6.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-80.

Valenzuela, J. G., Charlab, R., Gonzalez, E. C., Miranda-Santos, I. K. F., Marinotti, O., Francischetti, I. M. and Ribeiro, J. M. C. (2002a). The D7 family of salivary proteins in blood sucking Diptera. *Insect Mol. Biol.* **11**, 149-55.

Valenzuela, J. G., Francischetti, I. M., Pham, V. M., Garfield, M. K. and Ribeiro, J. M. (2003). Exploring the salivary gland transcriptome and proteome of the *Anopheles stephensi* mosquito. *Insect Biochem Mol Biol* **33**, 717-32.

Valenzuela, J. G., Francischetti, I. M. and Ribeiro, J. M. (1999). Purification, cloning, and synthesis of a novel salivary anti-thrombin from the mosquito *Anopheles albimanus*. *Biochemistry* **38**, 11209-11215.

Valenzuela, J. G., Pham, V. M., Garfield, M. K., Francischetti, I. M. and Ribeiro, J. M. C. (2002b). Toward a description of the sialome of the adult female mosquito *Aedes aegypti*. *Insect Biochem. Mol. Biol.* **32**, 1101-1122.

Vaughan, J. A., Noden, B. H. and Beier, J. C. (1991). Concentrations of human erythrocytes by anopheline mosquitoes (Diptera: Culicidae) during feeding. *J Med Entomol* **28**, 780-6.

Weintraub, N. L., Fang, X., Kaduce, T. L., VanRollins, M., Chatterjee, P. and

Spector, A. A. (1999). Epoxide hydrolases regulate epoxyeicosatrienoic acid incorporation into coronary endothelial phospholipids. *Am J Physiol* **277**, H2098-108.

Yamazaki, Y., Koike, H., Sugiyama, Y., Motoyoshi, K., Wada, T., Hishinuma, S., Mita, M. and Morita, T. (2002). Cloning and characterization of novel snake venom proteins that block smooth muscle contraction. *Eur J Biochem* **269**, 2708-15.

Yamazaki, Y. and Morita, T. (2004). Structure and function of snake venom cysteine-rich secretory proteins. *Toxicon* **44**, 227-31.

Zhu, Y. Y., Machleder, E. M., Chenchik, A., Li, R. and Siebert, P. D. (2001). Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. *Biotechniques* **30**, 892-7.

Zimmermann, B., Dames, P., Walz, B. and Baumann, O. (2003). Distribution and serotonin-induced activation of vacuolar-type H⁺-ATPase in the salivary glands of the blowfly *Calliphora vicina*. *J Exp Biol* **206**, 1867-76.

Figure legends

Fig. 1. Wolbachia-like genes of *Anopheles gambiae* in chromosome 3R. Blue boxes represent automatic annotation of putative genes. Dark grey boxes represent matching location of EST. Vertical dark lines represent stop codons on selected forward (top) or reverse (bottom) frame.

Fig. 2. Transposable elements (TE5P and TE3P) flanking the 5' and 3' Wolbachia gene region in *Anopheles gambiae* chromosome 3R.

Fig. 3. BLAST result analysis of transposable elements (TE5P and TE3P) flanking the Wolbachia gene region in chromosome 3R. (A) TEP5. (B) TEP6.

Fig. 4. Diagram of the D7 gene cluster on chromosome arm 3R. The arrow representing Contig_709 indicates a non-coding RNA mapping to the 5' region of the D7 short cassette.

Fig. 5. Diagram of the SG1 gene cluster on chromosome X of *Anopheles gambiae*. Blue bars represent genes, and dark bars represent RNA sequences found in the salivary transcriptome of adult females.

Fig 6. Clustal alignment of the gSG1/Trio family of salivary proteins.

Fig. 7. Tissue specificity of genes expressed in the salivary glands of adult female *Anopheles gambiae* mosquitoes as evidenced by RT-PCR analysis. Total RNA from female salivary glands (sg), carcasses (c; adult females with salivary glands removed), or adult males (m) was amplified using the gene-specific primers indicated on the right. Amplifications are grouped according to pattern of expression. (A) housekeeping; (B) female gland specific; (C) enriched in female glands; (D) female glands and adult males; (E) rpS7, ribosomal protein S7 used for normalization.

Fig 8. Tissue-specific alternative splicing of sal_tryp XII. Total RNA from sg, c, and m (see legend to Fig. 5) was amplified by RT-PCR. The length in base pair of the amplified fragments is indicated.

Figure 1: *Wolbachia*-like genes of *Anopheles gambiae* in chromosome 3R. Blue boxes represent automatic annotation of putative genes. Dark grey boxes represent matching location of EST's. Vertical dark lines represent stop codons on selected forward (Top) or reverse (Bottom) frame.



Figure 2: Transposable elements (TE5p and TE3p) flanking the 5' and 3' Bacterial *Wolbachia* gene region in *Anopheles gambiae* chromosome 3R.

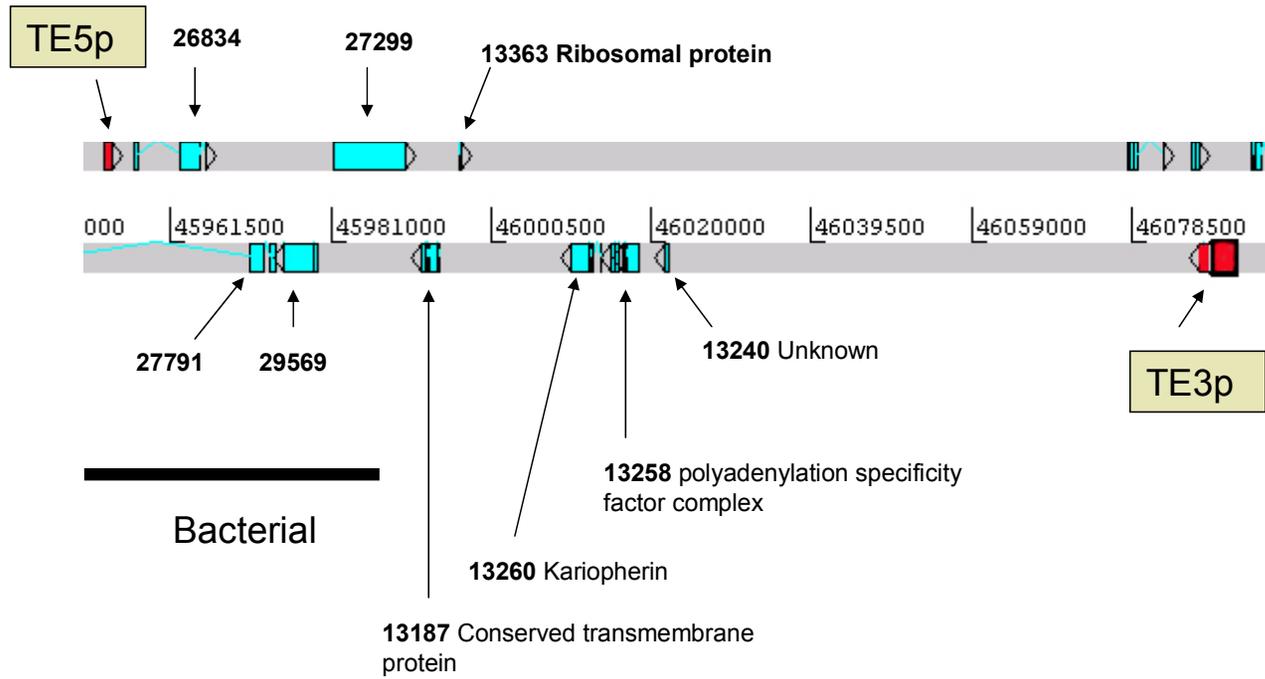


Figure 3: Blast result analysis of transposable elements (TE5P and TE3P) flanking the *Wolbachia* gene region in chromosome 3R. A) TEP5 B) TEP6.

A: Blast TEP5 comparison with *C. elegans* transposon

```

=> GI|17565388|REF|NP_503526.1| PREDICTED CDS, REVERSE TRANSCRIPTASE
      FAMILY MEMBER (5C378) (Handa and Guidotti, 1996)
      Length = 417

Score = 109 bits (272), Expect = 1e-024
Identities = 86/327 (26%), Positives = 151/327 (46%), Gaps = 22/327 (6%)
Frame = +1

Query: 37  TKGLRQGDGLACLLFNALERAIRD-SRVETTGT-----IFYKSTQILACADDI 180
          TKG+RQGD ++ LF+ LE R S+E G + K+ L ADDI
Sbjct: 96  TKGVRQGDPISEPNLFSACLEHVFRKLSCEIEFKGEAEDYNTIPGMRVNGKLNLTNLRFADDI 155

Query: 181  DIIGLRLSYVAEAYQGIEQEAESLGLQINEAKSKLMVATSAGLPINNQNLRRRDVQIGER 360
          +I + ++ Q + Q+ +GL+IN K+K++ N+ V G
Sbjct: 156  VLIANHPNTASKMLQELVQKCSEVGLLEINTGKTKVL-----RNRFA DPSKVYFGNP 206

Query: 361  T----FEVVPQFTYLGSKVSNDNSMEAE LRARMLAANRSFYSLKKQFTSKNLLRRTKLGL 528
          + + V ++ YLG +++ N++ E+ R AA +F +K T ++ + L
Sbjct: 207  SSTTQLDDVDEYIYLRQINAQNNLMPEIHRRRRAAWAAFNGIKNT-TDSITDKKIRANL 265

Query: 529  YSTYIVPVLTYASETWTWLFKSDKTLA AAFERKMLRRLIGPVCVEGQWRSRYNDELYEMYG 708
          + + ++P LTY SE WT K+ + + RR++G + + R + +++ +M
Sbjct: 266  FDSIVLPALTYGSEAWTFTKALSERVRITHASLERLRVGI TLTQQRERDLHREDIRKMSL 325

Query: 709  DLTVVQRIKLARLRWAGHVVRMETDDPARKVFLGRPQGQRRR-GRPCLRWDGVE----- 870
          + +K +L WAGHV R + + RP G +R GRP +RW D +
Sbjct: 326  VRDPLNFVKKRKLGWAGHVARRKDGRTTLMTEWRPYGWKRPVGRPPMRWTD SLRKEITT 385

Query: 871  ASAINAGITDWQTKARDRERFR TLLKQ 951
          A IT W T A+DR+ + ++++
Sbjct: 386  RDADGEVITPWSTIAKDRKEWLAVIRR 412
  
```

3B: Blast TEP3 comparison with TF2 Transposable element of *S. pombe*

```

=> GI|19075455|REF|NP_587955.1| TF2 TYPE TRANSPOSABLE ELEMENT
      (Tatusov et al., 2003)
      Length = 1333

Score = 194 bits (494), Expect = 4e-050
Identities = 149/556 (26%), Positives = 269/556 (48%), Gaps = 29/556 (5%)
Frame = +1

Query: 16  EAFETLREKLTQSPILAIFDPKKETELHCDASSFGFGAVLLQKQEDNKLHPVAYFSKTTS 195
          +A E +++ L P+L FD K+ L DAS GAVL QK +D+K +PV Y+S S
Sbjct: 684  QAIENIKQCLVSPVLRHFDFSKKILLETDASDVAVGAVLSQKHDDDKYYYPVGYYSKMS 743

Query: 196  KEEAKLHSELETL SVIYALKRFHVYVHGL--PLKIFTDCNSLVETLKNRN--ASAKIAR 363
          K + + E L++I +LK + Y+ P KI TD +L+ + N + + ++AR
Sbjct: 744  KAQLNYSVSDKEMLA I KSLKHWRHYLESTIEPFKILTDRNLIGRITNESEPENKRLAR 803

Query: 364  WSLFLENYDYTIHYRSGTSMHVDALSR----TEAVGAIS-DLDFLQQLQIAQSQD---- 516
          W LFL+++++ I+YR G++ DALSR TE + S D ++F QI+ + D
Sbjct: 804  WQLFLQDFNF EINYRPGSANHIADALSRIVDETEPIPKDSEDNSIN FVNQISITDDFKNQ 863

Query: 517  -----PLINTLRQKLEAGSVQGFILQDGLVYRQSSTNHLQLYVPREMVDNIIRHNN 669
          L+N L + + + L+DGL+ +S + + L ++ II+ H
Sbjct: 864  VVTEYNTDNTKLLNLLNNE-DKRVEENIQLKDGLLI--NSKDQILLPNDTQLTRTIIKKYH 920

Query: 670  E--KIGHLAISKTCQTISQHYWFPHPKPRVENFIKNC LKCI VYSAPPRTNRRNMYSIPKT 843
          E K+ H I I + + + ++ +++ ++NC C + + + IP +
Sbjct: 921  EEGKLIHPGIELLTNIILRRFTWKGIRKQIQEYVQNCHTCQINKSRNHKPYGPLQPIPPS 980
  
```

Query: 844 PVPFDTLHIDHLGPLPNITSRKKYILVVIDAFTKFKLYATATTNTNEVCEAL--TQYMS 1017
P+++L +D + LP +S + VV+D F+K L + T E + + ++
Sbjct: 981 ERPWESLSMDFITALPE-SSGYNALFVVVDRFSKMAILVPCTKSITAEQTARMFDQRVIA 1039

Query: 1018 YYSRPKRIISDRATCFSTSTAFKEFVDSNDITHVLNATGSPQANGQVERVNRVLRPILSKL 1197
Y+ PK II+D FTS +K+F + + PQ +GQ ER N+ + +L +
Sbjct: 1040 YFGNPKEIIADNDHIFTSQTWKDFAHKYNFVMKFSLPYRQTDGQTERTNQTVKLLRCV 1099

Query: 1198 CNSSDHSWSSHLRSAEHALNNTVHSSTNFLPSILLFGIEQRGQILDELHEFLNDKHVTT 1377
C S+ + W H+ + + NN +HS+T P ++ L EL F +DK
Sbjct: 1100 C-STHPNTWVDHISLVQQSYNNAIHSATQMTFFEIVHRYPALSPL-ELPSF-SDKTDEN 1156

Query: 1378 NRDLNLLRSEALSNIIEYSQYRNEQYVANRTKPAPSFSEGDLVAIKYTDS--TNANEKLV 1551
+++ + ++ + + ++Y + + F GDLV +K T + + + KL
Sbjct: 1157 SQETIQVFQTVKEHLNNTNNIKMKKYFDMKIQEIEEFQPGDLVMVKRRTKTGFLHKS NKLAP 1216

Query: 1552 KFRGP-YIVHKVLPD 1596
F GP Y++ K P++
Sbjct: 1217 SFAGPFYVLQKSGPNN 1232

Figure 4: Diagram of the D7 gene cluster on chromosome arm 3R. The arrow representing Contig_709 indicates a non-coding RNA mapping to the 5' region of the D7 short cassette.

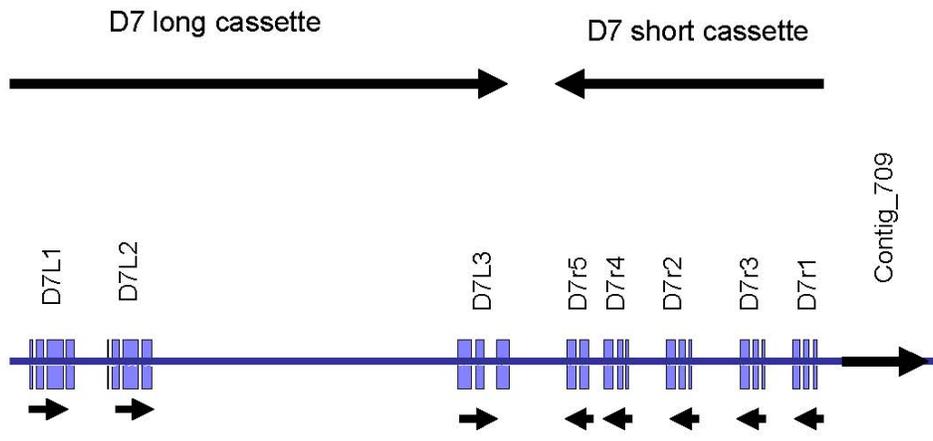


Figure 5: Diagram of the SG1 gene cluster on Chromosome X of *Anopheles gambiae*. Blue bars represent genes and dark bars represent RNA sequences found in the salivary transcriptome of adult females.

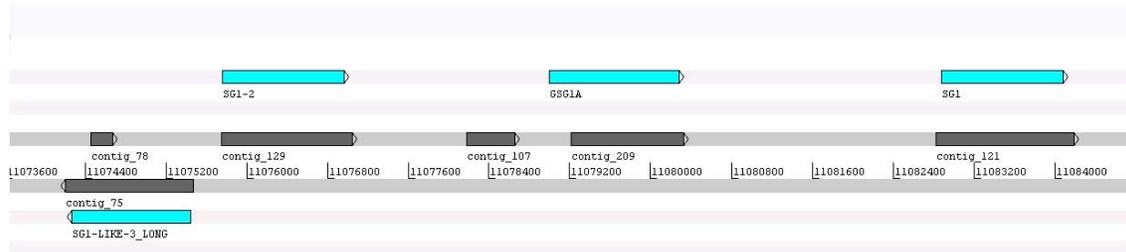


Figure 6: Clustal alignment of the gSG1/Trio family of proteins

```

GSG1A      -----ACTLQVPETMITRLLQAD-QPAGGCAAAWDSLLSRLEQMRHNLTCSE
SG1        -----EQCVIQLVGMVTRLLGPN-QAALSCDNLWSNLLLSFNFTRQDLVACQD
SG1-LIKE-3_LONG  RGLPESSDK-LEACGQHYGALLKASTTWNE--KE--CNGST-KLAA-----CVV
GSG1B     DGGSFLAPSGSNYCPIPLEVLEQENGTAPADWSAS-CTQRRTEDH-----AK
SG1-2     --QOTPEQ---SACVSHAGAILTAPLIHCHPLKA--CDGEAILKP-----
TRIO      ---EEAPKPEKEICGLKVGRLLDSVKGWLSVSQQEKCPLNKYCENK-----IQ
          *                               *

GSG1A     RETTNPAPPDDTRTNAPCQLLLNE---LQRKGNQ-TLLQLDKEMASKALYERIELAKVVQ
SG1       GELLVH-PPEP--YFSDCQRLDS---AKRDAEA-DRRAFTAEMQKKIQVNQWEADRYVQ
SG1-LIKE-3_LONG  SEHEQA-YRE---LKQRCQEAHDERT-AKVNAIYEKLPAYLSEVSARVNVLQVSLQHDLP
GSG1B     IEQAVA-----VIKDHLEKQA---ATKPIRDELRQFGGTLLPLLNSAQ--VKTDAA
SG1-2     --YLA-----LYQQCRINASERANLRDNFLF-RLNHWHDDHPFLETVAKAYEPLKA
TRIO      ADQYNL-----VPLTCIRWRSLNP-ASPTGSL-GGKDVSKIDAAMSNFKTLFEPMKA

GSG1A     DQDRIQQAIETMRTEQ----RNHRQLLLLYLDAADLRRALHQYK-LLVAQGDRHLPQQI
SG1       ESDTIRNQLTRLRDEL----RSTYRSLVLMSVOGGASKQALKYYHYLEGQPPGLLSSI
SG1-LIKE-3_LONG  NLQEMVGEQRHLIEQAWQYGAQLQHELMLTSMESDRVQRALVLHSMLVNAS-----LAEM
GSG1B     ELDELR-----MSVLVAAIEAGRIPEAMTQFLMLAGWN----RWFQI
SG1-2     ALESEQAHRSLRLVP-----ELQRKQLIYSIEAGREDAHILHMTLKGWK----PDQI
TRIO      DLAKLEEVKRQVLDAWKALEPLQEVYRSTLASGRIERAVFYSFMEMGDN-----
          .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
          .       .       .       .       .       .       .       .       .       .       .       .       .       .       .

GSG1A     VKFVYAAPRHENRRLENLLDLVRQLPARQDORTLYQLLQPELMKRPTQNQSTLAMLTALE
SG1       IAAVSVPEHAHERFEYLLDFARKVPGNSVRLAFYHQITAELRRHPEQRDSYLAMIALD
SG1-LIKE-3_LONG  VKESYQTHGADGRMVVRMLKFVRLLPGADERVAVYKQL-AELLKSNGQDGRFPAVIFSTD
GSG1B     VAQIYQNTRRDRQHIANLLEFIRIVPARDDRVAFYHELKKHVASKDYESYLGAMEADA
SG1-2     VSAIQDGYHVNPTIMEHLLEFVRAIPVRKERAAYKAIGPVIRNFKLTLTYVTLLEFAGDA
TRIO      ---VKLDNYFQPANVELLKYAWALPMHKKRSMYDLIGQLVQSSK---SPMLQTLHAVE
          .       .       .       .       .       .       .       .       .       .       .       .       .       .       .
          .       .       .       .       .       .       .       .       .       .       .       .       .       .       .

GSG1A     MG-----QVVEGNELKKQDAMYQLVLKRWMFLCLAGQ-YREIVQFATKHPRLFEQI
SG1       LGKL-----AFLSQNADARQLYLGLFNPALKRLQAAVIAGS-YDEIVTFATAPDHFEQI
SG1-LIKE-3_LONG  VR-----QLEDRYKPDHAQYEGKVVERWLAELQAGT-FHEVVEFARDYPEFARV
GSG1B     FHVVYEADGKTPLNESDVKALYTMLDGAGAYFQRALLTGANRYDLFLDEHHPQLFDLL
SG1-2     TG-----VFDTRKDRDDYISPLTVFVHMLRWQLANLEFEFHLSLAERFPRYSLH
TRIO      LA-----TVVNPELENRENLLNDQVVQLRDNLYKNS-FATLVSIARHFPDHFDTL
          .       .       .       .       .       .       .       .       .       .       .       .       .       .       .

GSG1A     GAKIATIHPKYWWKFS-FTQFVTYPNLLPLPEQRLEAFRTIMKQLQRNG-----KFFAD
SG1       ANQLSTLESEAWNQAN-FERLLLYPNRLPLAKQRLEAFRMLLQINQRNK-----HDFAD
SG1-LIKE-3_LONG  EEPLYETLKQWSAEG-LDRMVSFPNALPVGQRVRALRALLETLLQHQGE----QNNDV
GSG1B     FDRIVNVSQANMRKFN-SWQMMGALCRVHRPMSKVLLFRKTANLLLDHFKWE---KENEY
SG1-2     IEQIFFAPPHWQKAE-KRRLFEIASLFKAGHRFAAIEQLLKWAHQYAPMRGGKAHLEE
TRIO      RQRLFKLPDGSKPGADTLPNIVNFIAQLPSDELRLSSVDLLQSLTAENGT---LVQDPE
          :       .       .       .       .       .       .       .       .       .       .       .       .       .       .

GSG1A     HLAKLAPQIERCEQFLRQ-QKKELEWKDELVKLGQFADFDR-----RKDYAYLKNS-
SG1       RLTRVAKELDKCESFVKSGSKTQQSDLEKLVAVKRMFATRDA-----NRDYDHYLQES-
SG1-LIKE-3_LONG  YLIRLAHETGRVEATVQADAVRQALDDVKLFEQFKYQRG-----FPDYEALYKLF-
GSG1B     YAPMLAGYFEVCLPDIRK-DPATAGLVTEVQNIFGRYKKGMN-----YKSISHIIGKNI
SG1-2     ILPTLALETEKLRRTVEQTGKS-AELVRLKKLEEKFVSKDRWNTNTYRTYLHMIKTKG
TRIO      YVYRLSQLAHAMPSLDV--KAHPDLQSVDDLMAKFNTPIDGKTLQFQNIGISPSSSV
    
```

Figure 7: Tissue specificity of genes expressed in the salivary glands of adult female *Anopheles gambiae* mosquitoes as evidenced by Reverse Transcription PCR analysis. Total RNA from female salivary glands (sg), carcasses (c, adult females with salivary glands removed) or adult males (m) was amplified using the gene-specific primers indicated on the right. Amplifications are grouped according to pattern of expression: (A) housekeeping; (B) female gland-specific; (C) enriched in female glands; (D) female glands and adult males; (E) rpS7, ribosomal protein S7 used for normalization.

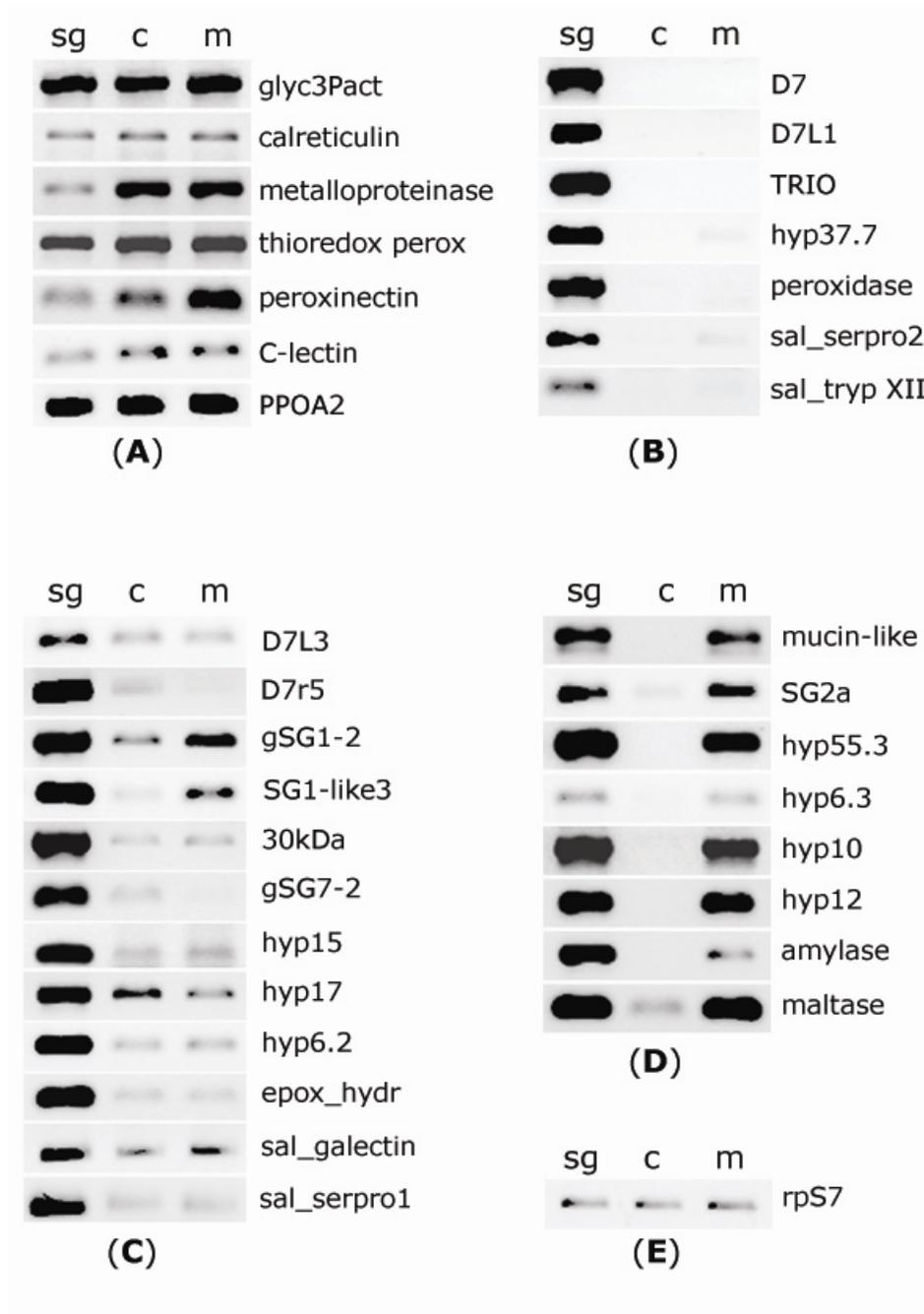


Figure 8: Tissue-specific alternative splicing of sal_tryp XII. Total RNA from sg, c and m (see legend to figure 7) was amplified by RT-PCR. The length in base pair of the amplified fragments is indicated.

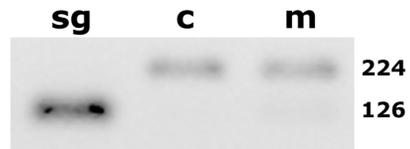


Table 1: Functional Classification of salivary gland transcriptome of adult female *Anopheles gambiae* mosquitoes, based on transcript similarity to gene or adjacent untranslated region (UTR)

Transcript class	Based on Proteome match		Based on Proteome + UTR match	
	Number of Contigs	Number of Sequences	Number of Contigs	Number of sequences
Housekeeping	357	554	548	762
Unknown	337	382	132	157
Secreted	155	1940	166	1954
Bacterial	9	32	11	34
TE	3	3	4	4

Table 2: Classification of housekeeping genes expressed in salivary glands of adult female *Anopheles gambiae* mosquitoes. For more information, see text.

Functional class	Number of Contigs	Number of Seqs	Seq/contig
protein synthesis machinery	94	218	2.32
signal transduction	69	81	1.17
conserved of unknown function	56	71	1.27
transporters	52	63	1.21
protein modification machinery	44	56	1.27
metabolism, energy	38	59	1.55
protein export machinery	38	45	1.18
transcription factors	32	33	1.03
nuclear metabolism and regulation	29	30	1.03
cytoskeletal	21	30	1.43
transcription machinery	19	20	1.05
proteasome machinery	16	16	1.00
metabolism, lipid	10	11	1.10
metabolism, carbohydrate	8	10	1.25
membrane proteins of unknown function	8	8	1.00
cell adhesion	7	7	1.00
extracellular matrix	6	9	1.50
nitrogen excretion metabolism	6	7	1.17
cofactor synthesis	5	6	1.20
metabolism, amino acid	4	4	1.00
metabolism, nucleic acid and nucleotides	3	3	1.00
metabolism, oxidant	2	2	1.00
metabolism, heme	1	1	1.00
protease inhibitor	1	1	1.00

Table 3: Transcription factors and regulators possibly expressed in adult female salivary glands of *Anopheles gambiae*

Protein	Transcription factor	Class	References
ENSANGP00000016160	XBP-1	chaperones expression in ER	(Lee et al., 2003)
ENSANGP00000010250	VRI (vrille) protein	circadian	(Blau and Young, 1999))
ENSANGP00000018277	Timeless protein	circadian	(Young, 1996)
ENSANGP00000011468	MBF1	general	(Takemaru et al., 1997)
ENSANGP00000014526	Repressor of E1A-stimulated genes CREG	general	(Veal et al., 1998)
ENSANGP00000016990	Purine-rich binding protein- α	general	(Oikawa and Yamada, 2003), 2003)
ENSANGP00000029048	ADF1	general	(Cutler et al., 1998)
ENSANGP00000017163	CtBP - interacts with hairy	salivary gland differentiation	(Poortinga et al., 1998))
ENSANGP00000018854	hairy/enhancer-of-split related	salivary gland differentiation	(Myat and Andrew, 2002)
ENSANGP00000020858	Transcriptional repressors of the hairy/E(spl)	salivary gland differentiation	(Bianchi-Frias et al., 2004)
ENSANGP00000019039	Forkhead/HNF-3-related transcription factor	salivary gland differentiation	(Mach et al., 1996; Myat and Andrew, 2000)
ENSANGP00000022150	Fork head protein	salivary gland differentiation	(Zhou et al., 2001)
ENSANGP00000020686	Homeobox extradenticle	salivary gland differentiation	(Myat and Andrew, 2000; Peers et al., 1995))
ENSANGP00000004060	doublesex	sex differentiation	(Baker et al., 1989)
ENSANGP00000012953	Sex comb on midleg	sex differentiation	(Bornemann et al., 1998)
ENSANGP00000019791	combgap homeobox	tissue differentiation	(Simon et al., 1993)
ENSANGP00000020897	Grunge	tissue differentiation	(Erkner et al., 2002)
ENSANGP00000024958	Split ends long isoform Wingless signaling	tissue differentiation	(Lin et al., 2003)
ENSANGP00000027998	Capicua	tissue differentiation	(Goff et al., 2001; Jimenez et al., 2000))
ENSANGP00000003712	Cap-n-collar	tissue differentiation	(Mohler et al., 1995; Walker et al., 2000))