The Journal of Experimental Biology 205, 2429–2451 (2002)
Printed in Great Britain © The Company of Biologists Limited 2002
JEB4209

Toward a catalog for the transcripts and proteins (sialome) from the salivary gland of the malaria vector *Anopheles gambiae*

Ivo M. B. Francischetti[1], Jesus G. Valenzuela[1], Van My Pham[1], Mark K. Garfield[2] and José M. C. Ribeiro[1,*]

[1]*Medical Entomology Section, Laboratory of Parasitic Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD20892-0425, USA* and [2]*Research Technology Branch, Twinbrook II, USA*

*Author for correspondence (e-mail: jribeiro@nih.gov)

Accepted 1 May 2002

Summary

Hundreds of *Anopheles gambiae* salivary gland cDNA library clones have been sequenced. A cluster analysis based on sequence similarity at e^{-60} grouped the 691 sequences into 251 different clusters that code for proteins with putative secretory, housekeeping, or unknown functions. Among the housekeeping cDNAs, we found sequences predicted to code for novel thioredoxin, tetraspanin, hemopexin, heat shock protein, and TRIO and MBF proteins. Among secreted cDNAs, we found 21 novel *A. gambiae* salivary sequences including those predicted to encode amylase, calreticulin, selenoprotein, mucin-like protein and 30-kDa allergen, in addition to antigen 5- and D7-related proteins, three novel salivary gland (SG)-like proteins and eight unique putative secreted proteins (Hypothetical Proteins, HP). The electronic version of this paper contains hyperlinks to FASTA-formatted files for each cluster with the best match to the nonredundant (NR) and conserved domain databases (CDD) in addition to CLUSTAL alignments of each cluster. The N terminus of 12 proteins (SG-1, SG-1-like 2, SG-6, HP 8, HP 9-like, 5′ nucleotidase, 30-kDa protein, antigen 5- and four D7-related proteins) has been identified by Edman degradation of PVDF-transferred, SDS/PAGE-separated salivary gland proteins. Therefore, we contribute to the generation of a catalog of *A. gambiae* salivary transcripts and proteins. These data are freely available and will eventually become an invaluable tool to study the role of salivary molecules in parasite-host/vector interactions.

Key words: *Anopheles gambiae*, mosquito, salivary gland, malaria, proteomics, transcriptome, genomics, blood-sucking insect.

Introduction

Malaria affects 200 million people worldwide and causes approximately 1.5 million deaths every year (Fauci, 2001). The disease is caused by *Plasmodium* parasites transmitted by the blood-sucking mosquito *A. gambiae*, known as the major vector of malaria in sub-Saharan Africa (Collins and Paskewitz, 1995). The parasite has a complex life cycle in the vector, where it becomes infective after invasion of the salivary gland (Cerami et al., 1992; Touray et al., 1992). *Plasmodium* is transmitted *via* the bite of an infected mosquito, which releases the sporozoite stage into the skin (Sidjanki and Vandeberg, 1997) together with saliva (Huribut, 1966). Saliva not only operates as a carrier to deliver the sporozoite into the host (Krettli and Miller, 2001), but also contains a number of pharmacologically active molecules which counteract host defenses that are triggered by blood feeding (Ribeiro, 1987). Although both genome surveys (Kappe et al., 2001; Carlton et al., 2001; Janssen et al., 2001) and genome projects (Gardner et al., 1998; Bowman et al., 1999) for *Plasmodium* spp. have been conducted, only more recently has systematic sequencing of *A. gambiae* genes been envisaged (Adam, 2001; Balter, 2001).

In an attempt to reveal the complexity of *A. gambiae* salivary glands, a high-throughput approach designed to identify a large number of cDNAs in the gland of this mosquito has been employed. Remarkably, only approximately 15 % of our cDNA isolates match *A. gambiae* sequences previously reported (Arcà et al., 1999); many of the remaining clusters have unknown functions. Generation of a set of *A. gambiae* salivary cDNAs, in addition to the *Plasmodium* genome currently available, may provide indispensable tools for the systematic and comprehensive analysis of molecules that may play an active role in the pathogenesis of malaria.

Materials and methods

Reagents

All water used was of 18 MΩ quality and was produced using a MilliQ apparatus (Millipore, Bedford, MA, USA).

Organic compounds were obtained from Sigma Chemical Corporation (St Louis, MO, USA) or as stated otherwise.

Mosquitoes

A. gambiae gambiae Giles mosquitoes were reared under the expert supervision of Mr André Laughinghouse. Insectary rooms were kept at 26±0.5 °C, with a relative humidity of 70 % and a 16 h:8 h light:dark photoperiod. Adult female mosquitoes used in the experiments were aged 0–7 days, took no blood meals, and were maintained on a diet of 10 % Karo syrup solution. Salivary glands from adult female mosquitoes were dissected and transferred to 20 µl Hepes saline (HS; NaCl 0.15 mol l^{-1}, 10 mmol l^{-1} Hepes, pH 7.0) in 1.5 ml polypropylene vials in groups of 20 pairs of glands in 20 µl of HS or as individual glands in 10 µl of HS. Salivary glands were kept at –75 °C until needed.

Salivary gland cDNA library construction

A. 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). The polymerase chain reaction (PCR)-based cDNA library was made following the instructions for the SMART cDNA library construction kit (Clontech, Palo Alto, CA, USA). *A. gambiae* salivary gland mRNA (200 ng) was reverse transcribed to cDNA using Superscript II Rnase H-reverse transcriptase (Gibco-BRL, Gaithersburg, MD, USA) and the CDS III/3′ PCR primer (Clontech) for 1 h at 42 °C. Second-strand synthesis was performed through a PCR-based protocol using the SMART III primer (Clontech) as the sense primer and the CDS III/3′ primer as antisense primer. These two primers create *Sfi*IA and *Sfi*IB sites at the ends of the nascent cDNA. Double-strand cDNA synthesis was done on a Perkin Elmer 9700 Thermal cycler (Perkin Elmer Corp., Foster City, CA, USA) using Advantage Klen-*Taq* DNA polymerase (Clontech). PCR conditions were the following: 94 °C for 2 min; 19 cycles of 94 °C for 10 s and 68 °C for 6 min. Double-strand cDNA was immediately treated with proteinase K (0.8 µg $µl^{-1}$) for 20 min at 45 °C and washed three times with water using Amicon filters with a 100 kDa cutoff (Millipore). Double-strand cDNA was then digested with *Sfi*I for 2 h at 50 °C. The cDNA was then fractionated using columns provided by the manufacturer (Clontech). Fractions containing cDNA of more than 400 base pairs (bp) in size were pooled, concentrated, and washed three times with water using an Amicon filter with a 100 kDa cutoff. The cDNA was concentrated to a volume of 7 µl. The concentrated cDNA was then ligated into a Lambda TriplEx2 vector (Clontech), and the resulting ligation reaction was packed using Gigapack Gold III from Stratagene/Biocrest (Cedar Creek, TN, USA) following the manufacturer's specifications. The obtained library was plated by infecting log-phase XL1-Blue cells (Clontech) and the amount of recombinants were determined by PCR using vector primers flanking the inserted cDNA and visualized on a 1.1 % agarose gel with ethidium bromide (1.5 µg ml^{-1}).

Sequencing of A. gambiae *cDNA library*

The *A. gambiae* salivary gland cDNA library was plated to approximately 200 plaques per Petri dish (150 mm diameter). The plaques were randomly selected and transferred to a 96-well polypropylene plate containing 100 µl of water per well. The plate was covered and placed on a gyratory shaker for 1 h at room temperature. The phage sample (5 µl) was used as a template for a PCR reaction to amplify random cDNA. The primers used for this reaction were sequences from the TriplEx2 vector and were named PT2F1 (5′-AAG TAC TCT AGC AAT TGT GAG C-3′), which is positioned upstream from the cDNA of interest (5′ end), and PT2R1 (5′-CTC TTC GCT ATT ACG CCA GCT G-3′), which is positioned downstream from the cDNA of interest (3′ end). Platinum *Taq* polymerase (Gibco-BRL) was used for these reactions. Amplification conditions were: 1 hold at 75 °C for 3 min, 1 hold at 94 °C for 2 min, and 33 cycles at 94 °C for 1 min, 49 °C for 1 min, and 72 °C for 1 min and 20 s. Amplified products were visualized on a 1.1 % agarose gel with ethidium bromide. The concentration of double-stranded cDNA was measured using Hoechst dye 33258 on a Flurolite 1000 plate fluorometer (Dynatech Laboratories, Chantilly, VA, USA). PCR reactions (3–4 µl) containing between 100 and 200 ng of DNA were then treated with exonuclease I (0.5 units $µl^{-1}$) and shrimp alkaline phosphatase (0.1 units $µl^{-1}$) for 15 min at 37 °C and 15 min at 80 °C on a 96-well PCR plate. This mixture was used as a template for a cycle-sequencing reaction using the DTCS labeling kit from Beckman Coulter Inc. (Fullerton, CA, USA). The primer used for sequencing (PT2F3) is upstream from the inserted cDNA and downstream from primer PT2F1. The sequencing reaction was performed on a Perkin Elmer 9700 thermocycler. Conditions were 75 °C for 2 min, 94 °C for 4 min, and 30 cycles of 96 °C for 20 s, 50 °C for 20 s and 60 °C for 4 min. After cycle-sequencing the samples, a cleaning step was done using the multiscreen 96-well plate cleaning system (Millipore). The 96-well multiscreening plate was prepared by adding a fixed amount (manufacturer's specification) of Sephadex-50 (Amersham Pharmacia Biotech, Piscataway, NJ, USA) and 300 µl of deionized water. After 1 h of incubation at room temperature, the water was removed from the multiscreen plate by centrifugation at 750 ***g*** for 5 min. After partially drying the Sephadex in the multiscreen plate, the whole cycle-sequencing reaction was added to the center of each well, centrifuged at 750 ***g*** for 5 min, and the clean sample was collected on a sequencing microtiter plate (Beckman Coulter Inc.). The plate was then dried on a Speed-Vac SC 110 model with a microtiter plate holder (Savant Instruments Inc, Holbrook, NY, USA). The dried samples were immediately resuspended with 25 µl of deionized ultrapure formamide (J. T. Baker, Phillipsburg, NJ, USA), and one drop of mineral oil was added to the top of each sample. Samples were either sequenced immediately on a CEQ 2000 DNA sequencing instrument (Beckman Coulter Inc.) or stored at –30 °C.

Sequence information cleaning

Raw sequences originating from the DNA sequencer were

assigned one of five letters in their result: ATCG for identified nucleotide bases, and N when the sequencer program could not call a base. Usually the beginning and ends of the sequences have a higher proportion of N calls. Sequences also contain primer and vector sequences used in library construction. For this reason, raw sequences were treated by a program written in VisualBasic 6.0 (VB) (Microsoft Corp., Redmond, WA, USA) as follows. (i) Sequences were analyzed in their first 80 bp for groups of four Ns, and, if found, the block of four Ns closer to position 80 was used to trim the raw sequence from this 5′ N-rich region. (ii) For sequences longer than 110 bp, windows of 10 bp were screened for the occurrence of four or more Ns above position 100. The positive window with the smallest position value was used to trim the sequence from the 3′ N-rich region. Sequences thus trimmed and having more than 10 % N content were discarded. (iii) Good quality and trimmed sequences were then searched for occurrence of the primers used in library construction (the SMART III primer as well as the CDS/R primers). A moving window the size of the primer was searched on the sequence for matches with the primer sequence. If more than a 70 % match was obtained, or if a contiguous match longer than 50 % of the length of the primer was observed, the sequence was trimmed at the beginning or end of the window, depending on the expected position of the primer. This simple algorithm avoided errors due to spurious insertions. (iv) The trimmed sequence was 'polished' by removing any trailing N residues. The sequence final N content was assessed, as well as its AT content and length. The final sequence was written to a FASTA-format file containing in its definition line the actions taken by the program.

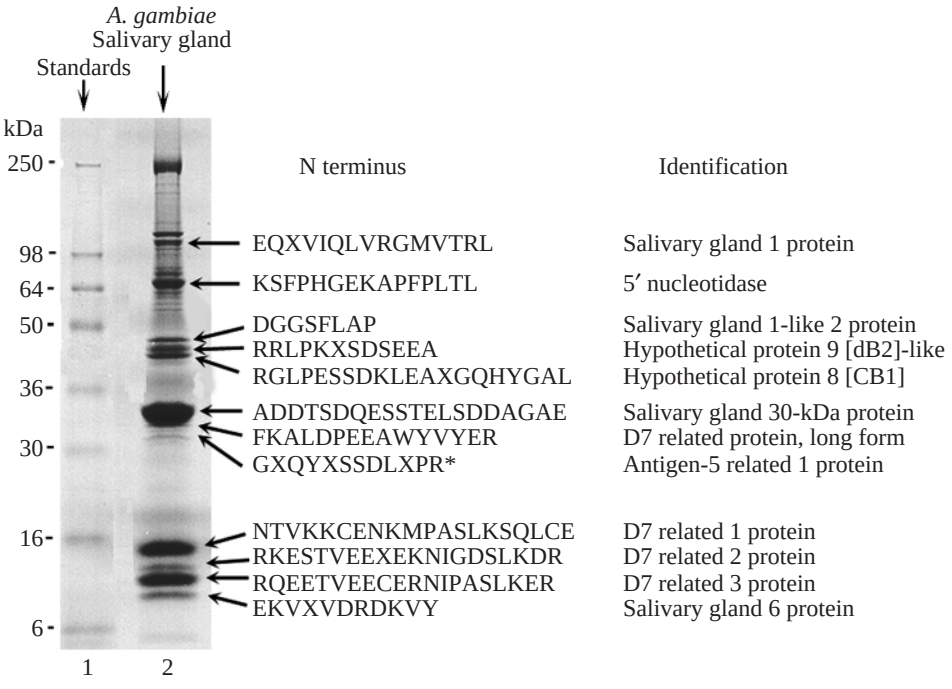

Fig. 1. SDS-PAGE of *A. gambiae* salivary proteins, under denaturing non-reducing conditions. Molecular mass markers are shown on the left. The match found is shown on the right. *Sequence was obtained using salivary protein separated by a 12 % NU-PAGE gel, Mops buffer, under denaturing non-reducing conditions.

Searches for known sequence similarities and known protein domains of the cDNA sequences

To obtain information on the possible role of the cDNA sequences, the FASTA file containing all the stripped sequences was blasted against the GenBank nonredundant protein database (NR) from the National Center for Biotechnology Information (NCBI) using the standalone BlastX program found in the executable package at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/ (Altschul et al., 1997). The NR database as well as the cumulative updates were regularly downloaded, uncompressed with GUNZIP (found at www.gzip.org/), and formatted for Blast program use with the FORMATDB program (executables also found at ftp://ftp.ncbi.nlm.nih.gov/blast/executables/) with the help of a program written in PERL code (software found at www.activeperl.com). NCBI sequences are indicated in this manuscript by their accession number as gi|XXXX where XXXX is a unique identifier number. The resulting file was parsed, and the best match was incorporated in the FASTA definition line after the delimiter |. The sequences were next submitted to the standalone program RPSBlast (Altschul et al., 1997) and searched against the Conserved Domains Database (CDD) (found at ftp://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/), which includes all Pfam (Bateman et al., 2000) and SMART (Schultz et al., 2000) protein domains. The RPSBlast result file was parsed as above and the best match incorporated also into the FASTA definition line of the sequence. When all sequences of a particular cluster were blasted against the NR protein database (using the BlastX program), the best protein match was searched for the species from which the NR database sequence originated. If the species was not *A. gambiae*, or no matches to the NR were found, the cluster was marked as representing a novel *A. gambiae* sequence (indicated by Y under the column marked N (novel) in Table 1). All cluster sequences that gave a match to an *A. gambiae* protein sequence were further individually inspected to verify whether the cDNA sequence represented nearly the same information translated as the protein match or a closely related but different protein. In this latter case, a Y would also be added to the results in Table 1 in the N (novel) column for the row of the cluster in question.

Sequence clustering

The FASTA file containing all sequences was clustered by first blasting (using the BlastN program) each sequence against the formatted database file using a Blast cutoff score of $1E^{-60}$. The resulting file was used to join in a single cluster all sequences that shared at least one common sequence in the BlastN result. Thus, if sequence A had a $1E^{-60}$ match to B and B had a similar match to C, the three sequences would be joined even if A had a less meaningful score in relation to C. The clustering program also made individual FASTA-formatted files for each cluster, sorted in descending order of sequence size. When these files contained two or more sequences, they were used as input for the sequence alignment program CLUSTALW (Higgins et al., 1996), which was called automatically by the clustering program. CLUSTAL alignment files were thus created for each cluster having two or more sequences. This clustering program was also written in VB. Finally, a program was written in VB that combines all the results to create Table 1 of this paper, except for the Function column. The output of this program is imported into a Microsoft Excel spreadsheet. In the supplemental material, Table 1 includes hyperlinks to the best NR protein match in the NCBI site, all FASTA files for each individual cluster, CLUSTAL alignment files for each cluster, when available, and the FASTA file for the whole database. Each cluster was individually analyzed for the probable function of its translation product and assigned a 'probably secreted', 'probably housekeeping' or 'indeterminate' function. This decision was based on the best match to the NR protein database and related sequences as searched online at the NCBI site (www.ncbi.nlm.gov) and on the SMART and/or Pfam matches, including searches of the nature of the domains by online searches of the respective sites.

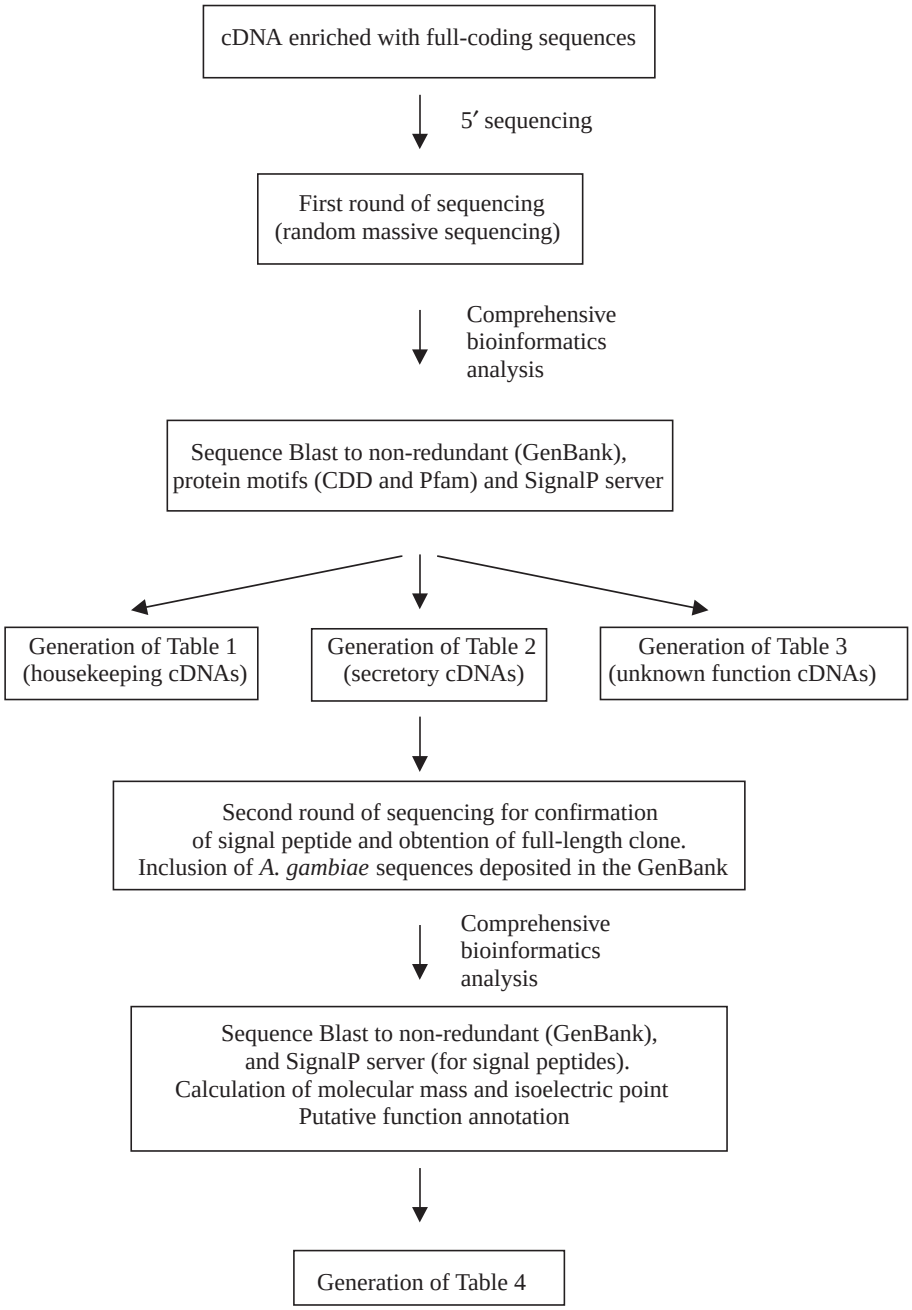

Fig. 2. Comprehensive bio-informatic analysis of 5′ EST from *A. gambiae* cDNA library and data generation. *A. gambiae* cDNA libraries enriched for full coding clones were constructed. 5′ EST sequences were systematically generated from the clones and analyzed for hits in nonredundant databases. Signal peptides were predicted by submission of the cluster sequences to the SignalP server, allowing the identification of putative secretory (S), housekeeping (H) or unknown (U) cDNAs. Putative functions were annotated according to database hits, when available.

Full-length sequencing of selected cDNA clones

A portion (4 μl) of the lambda phage containing the cDNA of interest was amplified using the PT2F1 and PT2R1 primers (conditions as described above). The PCR samples were cleaned using the multiscreen-PCR 96-well filtration system (Millipore). Cleaned samples were sequenced first with PT2F3 primer and subsequently with custom primers. Primer selection for complete sequence of selected full-length cDNA was also assisted by a program (written in VB) that identified unique primer sites within the sequences. To assemble the sequences,

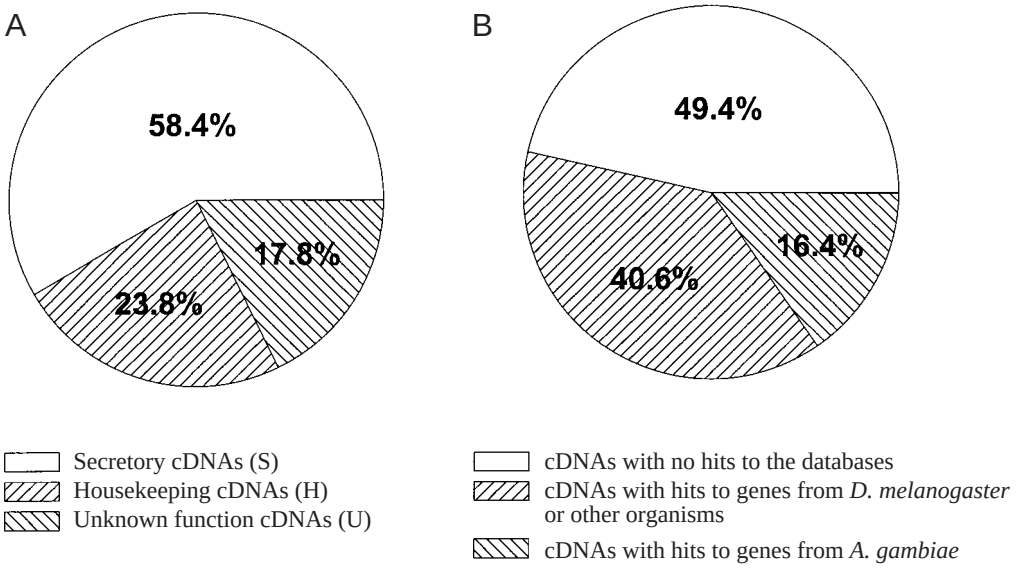

Fig. 3. *A. gambiae* salivary gland EST. The 691 individual sequences were assigned as secretory (S), housekeeping (H) or unknown (U) function cDNAs and the % representation is shown in (A). The 251 clusters were assigned as having no matches to the database, or having hits to sequences obtained from *D. melanogaster* (or other organisms), or from *A. gambiae*, and the % representation is shown in (B).

the previously known sequence was blasted against the new sequence using the standalone program bl2seq found with the executable package provided at the NCBI ftp site mentioned above. After identifying the regions of overlap, the two sequences were joined. The program attempted to locate a poly(A) region by using a 12-bp window in which 11 A residues would constitute a poly(A) string. If no poly(A) was found, a new set of primers would be found to continue extension of the cDNA. The program also generates CLUSTAL alignments of all sequences and produces a consensus output and the three possible translations of this unidirectionally cloned RNA. The final alignment is adjusted by hand. If necessary, the original tracings of the DNA sequencer are reviewed for critical base calls. The translated sequences are submitted as a FASTA file to the SIGNALP server (at http://www.cbs.dtu.dk/ services/SignalP/) (Nielsen et al., 1997), which responds by e-mail: indicating whether a signal peptide exists and its location. A program written in VB interprets this SIGNALP result file and removes the signal peptide, if it is predicted to exist, to create a mature protein sequence. Molecular masses using average molecular masses for C, H, O, N, P and S are calculated for all protein sequences, as are pI based on reduced proteins, following the pKa for amino acids within proteins as indicated before (Altland, 1990; Bjellqvist et al., 1994). This program, combined with the program generating Table 1 of this paper, produced an output that can be read by the spreadsheet program Excel to produce Table 2 in this paper. In the supplemental material, available on request, hyperlinks are given to all proteins.

SDS-PAGE

A precast 16 % polyacrylamide gel was used and run in Tris-glycine-SDS buffer. Alternatively, a NU-PAGE 12 % Bis-Tris gel, 1 mm thick (Invitrogen), was used and run in MOPS buffer according to the manufacturer's instructions. To estimate the molecular mass of the samples, SeeBlue™ markers from Invitrogen (myosin, bovine serum albumin, glutamic dehydrogenase, alcohol dehydrogenase, carbonic anhydrase, myoglobin, lysozyme, aprotinin and insulin, chain B) were used. Salivary gland homogenates were treated with SDS (2 %) or NU-PAGE LDS sample buffer (Invitrogen) without reducing conditions. 20 pairs of homogenized salivary glands per lane (approximately 20 µg protein) were applied when visualization of the protein bands stained with Coomassie Blue was required. For amino-terminal sequencing of the salivary proteins, 20 homogenized pairs of glands were electrophoresed and transferred to polyvinylidene difluoride (PVDF) membrane using 10 mM CAPS, pH 11.0, 10 % methanol as the transfer buffer on a Blot-Module for the Xcell II Mini-Cell (Invitrogen). The membrane was stained with Coomassie Blue in the absence of acetic acid. Stained bands were cut from the PVDF membrane and subjected to Edman degradation using a Procise sequencer (Perkin-Elmer Corp.). 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 VB) that checked these amino acid sequences against the three possible protein translations of each cDNA sequence obtained in the mass sequencing project. This program was written using the same approach utilized in the BLOCKS (Henikoff and Henikoff, 1994) or PROSITE (Sibbald et al., 1991) databases. The program is very useful when mixed sequence information occurs, for example, amino-terminal sequences deriving from a mix of equal peptides. In this case, two different cDNA sequences may be unambiguously found as matches.

Statistical tests

Statistical tests were performed with SigmaStat version 2.0 (Jandel Software, San Rafael, CA, USA). Kruskal–Wallis ANOVA on ranks was performed, and multiple comparisons were done by the Dunn method. Dual comparisons were made with the Mann–Whitney rank sum test.

Table 1. Anopheles gambiae *housekeeping cDNAs*

Clus[1]	R[2]	GenBank match[3]	E-value[4]	Gi numb[5]	PFAM match[6]	E-value[7]
43	2	gi|7294669| CG10361 gene product	6.00E-83	gi|7294669	pfam00222 aminotran_2	5.00E-49
164	1	gi|7299845| CG3172 gene product	3.00E-65	gi|7299845	Smart smart00102 ADF	1.00E-17
223	1	gi|7291724| tsr gene product	2.00E-46	gi|7291724	pfam00241 cofilin_ADF	7.00E-20
146	1	gi|14572580| arrestin [Anopheles gamb...	3.00E-37	gi|14572580	pfam02752 arrestin_C	3.00E-19
73	1	gi|7300478| ATPsyn-d gene product	4.00E-24	gi|7300478	No matches found	
179	1	gi|7294882| ATPsyn-b gene product	5.00E-38	gi|7294882	No matches found	
60	2	gi|3065729| clathrin light chain	3.00E-29	gi|3065729	pfam01086 Clathrin_lg_ch	3.00E-26
181	1	gi|7297903| CG6766 gene product	2.00E-23	gi|7297903	pfam00668 Condensation	0.001
33	2	gi|309068| cytochrome b [Anopheles	1.00E-53	gi|309068	pfam00033 cytochrome_b_N	2.00E-43
18	6	gi|309062| cytochrome c oxidase subunit	5.00E-70	gi|309062	pfam00510 COX3	5.00E-41
24	4	gi|309058| cytochrome c oxidase subunit	2.00E-70	gi|309058	pfam00115 COX1	2.00E-83
87	1	gi|309058| cytochrome c oxidase subunit	2.00E-41	gi|309058	pfam00115 COX1	1.00E-30
147	1	gi|309058| cytochrome c oxidase subunit	5.00E-11	gi|309058	No matches found	
195	1	gi|1946624| cytochrome c oxidase	2.00E-84	gi|1946624	pfam00116 COX2	1.00E-45
216	1	gi|7297373| CG7870 gene product	1.00E-62	gi|7297373	pfam00535 Glycos_transf_2	4.00E-18
36	2	gi|12667408| elongation fact...	6.00E-19	gi|12667408	pfam00679 EFG_C	4.00E-12
84	1	gi|12328436| elongation factor 1 delt...	1.00E-34	gi|12328436	LOAD_EF1B EF1B	1.00E-30
226	1	gi|12667408| elongation fact...	1.00E-89	gi|12667408	pfam00679 EFG_C	2.00E-26
239	1	gi|7297950| Elf gene product	1.00E-103	gi|7297950	pfam00009 GTP_EFTU	6.00E-19
65	1	gi|7303599| BcDNA:GM12291 gene	2.00E-22	gi|7303599	No matches found	
249	1	gi|7295381| CG10733 gene product	3.00E-21	gi|7295381	pfam01105 EMP24_GP25L	7.00E-09
74	1	gi|7303508| CG8860 gene product	3.00E-20	gi|7303508	pfam00584 SecE	2.00E-14
123	1	gi|7296662| CG1213 gene product [alt	4.00E-24	gi|7296662	pfam00083 sugar_tr	5.00E-06
116	1	gi|7301644| BcDNA:GH07066 gene	2.00E-21	gi|7301644	No matches found	
103	1	gi|7296006| GlyP gene product	2.00E-57	gi|7296006	pfam00343 phosphorylase	3.00E-38
70	1	gi|7293608| CG14207 gene product	2.00E-44	gi|7293608	pfam00011 HSP20	6.00E-20
235	1	gi|7291835| CG4859 gene product	2.00E-70	gi|7291835	pfam00045 hemopexin	1.00E-11
106	1	gi|7301711| CG14525 gene product	1.00E-29	gi|7301711	No matches found	
112	1	gi|7293724| CG7770 gene product	1.00E-25	gi|7293724	pfam01920 KE2	4.00E-21
61	1	gi|7291246| mago gene product	2.00E-55	gi|7291246	pfam02792 Mago_nashi	2.00E-44
79	1	gi|7299378| CG4591 gene product	0.007	gi|7299378	No matches found	
99	1	gi|9624383| tetraspanin E118...	1.00E-16	gi|9624383	pfam00335 transmembrane4	6.00E-39
30	3	gi|309061| adenosine triphosphatase	3.00E-62	gi|309061	pfam00119 ATP-synt_A	1.00E-28
48	2	gi|7297772| porin gene product	1.00E-67	gi|7297772	pfam01459 Euk_porin	5.00E-60
50	2	gi|1438862| ADP/ATP carrier protein	1.00E-101	gi|1438862	pfam00153 mito_carr	9.00E-25
3	38	gi|10434098| unnamed protein product...	0.002	gi|10434098	pfam02414 Borrelia_orfA	0.002
228	1	gi|7297173| Nrv2 gene product [alt 2]	5.00E-12	gi|7297173	pfam00287 Na_K-ATPase	8.00E-05
121	1	gi|3037018| NADH dehydrogenase	2.00E-05	gi|3037018	pfam01604 7tm_5	0.002
169	1	gi|309069| NADH dehydrogenase subunit 1	1.00E-83	gi|309069	pfam00146 NADHdh	2.00E-71
199	1	gi|12882| NADH dehydrogenase subunit 5	0.1	gi|12882	pfam00001 7tm_1	6.00E-04
51	2	gi|1655708| NM23/nucleoside	8.00E-48	gi|1655708	pfam00334 NDK	3.00E-31
201	1	gi|2708713| ornithine decarboxylase	4.00E-08	gi|2708713	pfam02100 ODC_AZ	8.00E-06
120	1	gi|7296285| CG5605 gene product [alt	2.00E-85	gi|7296285	pfam01605 RF1	2.00E-20
109	1	gi|7303822| CG1418 gene product	2.00E-45	gi|7303822	No matches found	
182	1	gi|7299121| Prosbeta3 gene product	4.00E-96	gi|7299121	pfam00227 proteasome	3.00E-27
104	1	gi|4115422| protein disulphide isomer...	4.00E-06	gi|4115422	No matches found	
42	2	gi|7302715| CG6370 gene product	2.00E-04	gi|7302715	No matches found	
23	4	gi|5690416| ribosomal protei...	1.00E-32	gi|5690416	pfam00181 Ribosomal_L2	4.00E-25
28	3	gi|7298519| CG10652 gene product	3.00E-47	gi|7298519	pfam01248 Ribosomal_L7Ae	1.00E-21
38	2	gi|10242304| ribosomal protein S26	1.00E-59	gi|10242304	pfam01283 Ribosomal_S26e	4.00E-56
40	2	gi|7294664| CG7283 gene product	3.00E-69	gi|7294664	pfam00687 Ribosomal_L1	4.00E-35
45	2	gi|4239711| ribosomal protein P2	9.00E-23	gi|4239711	pfam00428 60s_ribosomal	3.00E-34
49	2	gi|7290609| CG4111 gene product	3.00E-30	gi|7290609	pfam00831 Ribosomal_L29	4.00E-13

Continued on p. 2436.

Table 1. *Continued*

Org[8]	N[9]	Function[10]	F[11]	Clone[12]	FASTA[13]	CLUSTAL[14]
Drosophila melanogaster	Y	2-amino-3ketobutyrate-Coa Ligase	H	NOSIG	AGN10E60-43clu.txt	AGN10E60-43aln.txt
Drosophila melanogaster	Y	Actin depolarization factor	H	NoORF	AGN10E60-164clu.txt	
Drosophila melanogaster	Y	Actin depolarization factor	H	NoORF	AGN10E60-223clu.txt	
Anopheles gambiae	N	Arrestin	H	NoORF	AGN10E60-146clu.txt	
Drosophila melanogaster	Y	ATP synthase	H	NoORF	AGN10E60-73clu.txt	
Drosophila melanogaster	Y	ATPsynthase	H	ANCH	AGN10E60-179clu.txt	
Drosophila melanogaster	Y	Clathrin chain	H	NoORF	AGN10E60-60clu.txt	AGN10E60-60aln.txt
Drosophila melanogaster	Y	Condensation domain used in non-ribosomal peptide synthesis	H	NoORF	AGN10E60-181clu.txt	
Anopheles gambiae	N	Cytochrome b	H	ANCH	AGN10E60-33clu.txt	AGN10E60-33aln.txt
Anopheles gambiae	N	Cytochrome oxidase	H	NOSIG	AGN10E60-18clu.txt	AGN10E60-18aln.txt
Anopheles gambiae	N	Cytochrome oxidase	H	NoORF	AGN10E60-24clu.txt	AGN10E60-24aln.txt
Anopheles gambiae	N	Cytochrome oxidase	H	NoORF	AGN10E60-87clu.txt	
Anopheles gambiae	N	Cytochrome oxidase	H	NOSIG	AGN10E60-147clu.txt	
Anopheles gambiae	Y	Cytochrome oxidase	H	NoORF	AGN10E60-195clu.txt	
Drosophila melanogaster	Y	Dolichyl-phosphate beta-glucosyltransferase	H	NoORF	AGN10E60-216clu.txt	
Anopheles gambiae	Y	Elongation factor	H	NoORF	AGN10E60-36clu.txt	AGN10E60-36aln.txt
Bombyx mori	Y	Elongation factor	H	NoORF	AGN10E60-84clu.txt	
Anopheles gambiae	Y	Elongation factor	H	NOSIG	AGN10E60-226clu.txt	
Drosophila melanogaster	Y	Elongation factor	H	NOSIG	AGN10E60-239clu.txt	
Drosophila melanogaster	Y	ER protein	H	NoORF	AGN10E60-65clu.txt	
Drosophila melanogaster	Y	ER protein	H	NOSIG	AGN10E60-249clu.txt	
Drosophila melanogaster	Y	ER translocation protein	H	ANCH	AGN10E60-74clu.txt	
Drosophila melanogaster	Y	Glucose transporter	H	NoORF	AGN10E60-123clu.txt	
Drosophila melanogaster	Y	Glycerol phosphate acyltransferase	H	NoORF	AGN10E60-116clu.txt	
Drosophila melanogaster	Y	Glycogen phosphorylase	H	NOSIG	AGN10E60-103clu.txt	
Drosophila melanogaster	Y	Heat shock protein	H	NoORF	AGN10E60-70clu.txt	
Drosophila melanogaster	Y	Hemopexin	H	NOSIG	AGN10E60-235clu.txt	
Drosophila melanogaster	Y	Highly conserved among species	H	ANCH	AGN10E60-106clu.txt	
Drosophila melanogaster	Y	KE2 family	H	NoORF	AGN10E60-112clu.txt	
Drosophila melanogaster	Y	Mago nashi protein –salivary gland differentiation?	H	NoORF	AGN10E60-61clu.txt	
Drosophila melanogaster	Y	Membrane protein?	H	NoORF	AGN10E60-79clu.txt	
Manduca sexta	Y	Membrane protein?	H	NoORF	AGN10E60-99clu.txt	
Anopheles gambiae	N	Mitochondrial enzyme	H	NoORF	AGN10E60-30clu.txt	AGN10E60-30aln.txt
Drosophila melanogaster	Y	Mitochondrial protein	H	NoORF	AGN10E60-48clu.txt	AGN10E60-48aln.txt
Anopheles gambiae	N	Mitochondrial protein	H	NoORF	AGN10E60-50clu.txt	AGN10E60-50aln.txt
Homo sapiens	Y	Mitochondrion genome gi|5834911	H	NoORF	AGN10E60-3clu.txt	AGN10E60-3aln.txt
Drosophila melanogaster	Y	Na+/K+ ATPase subunit	H	NoORF	AGN10E60-228clu.txt	
Bodo saltans	Y	NADH dehydrogenase	H	NoORF	AGN10E60-121clu.txt	
Anopheles gambiae	N	NADH dehydrogenase	H	NoORF	AGN10E60-169clu.txt	
Crithidia oncopelti	Y	NADH dehydrogenase	H	NoORF	AGN10E60-199clu.txt	
Xenopus laevis	Y	Nucleoside diphosphate kinase	H	NOSIG	AGN10E60-51clu.txt	AGN10E60-51aln.txt
Drosophila melanogaster	Y	Ornithine decarboxylase	H	NOSIG	AGN10E60-201clu.txt	
Drosophila melanogaster	Y	Peptide chain relase factor	H	NoORF	AGN10E60-120clu.txt	
Drosophila melanogaster	Y	Prenylated rab acceptor	H	NOSIG	AGN10E60-109clu.txt	
Drosophila melanogaster	Y	Proteasome protein	H	NOSIG	AGN10E60-182clu.txt	
Antheraea pernyi	Y	Protein disulphide isomerase?	H	NOSIG	AGN10E60-104clu.txt	
Drosophila melanogaster	Y	Ribophorin	H	NoORF	AGN10E60-42clu.txt	AGN10E60-42aln.txt
Anopheles gambiae	N	Ribosomal protein	H	NoORF	AGN10E60-23clu.txt	AGN10E60-23aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-28clu.txt	AGN10E60-28aln.txt
Anopheles gambiae	N	Ribosomal protein	H	NOSIG	AGN10E60-38clu.txt	AGN10E60-38aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-40clu.txt	AGN10E60-40aln.txt
Ceratitis capitata	Y	Ribosomal protein	H	NO SIG	AGN10E60-45clu.txt	AGN10E60-45aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NoORF	AGN10E60-49clu.txt	AGN10E60-49aln.txt

Table 1. *Continued*

Clus[1]	R[2]	GenBank match[3]	E-value[4]	Gi numb[5]	PFAM match[6]	E-value[7]
52	2	gi|4585827| ribosome associated membr...	4.00E-11	gi|4585827	No matches found	
53	2	gi|7301299| CG10423 gene product	3.00E-40	gi|7301299	pfam01667 Ribosomal_S27e	6.00E-21
54	2	gi|9438108| putative large s...	2.00E-45	gi|9438108	pfam01199 Ribosomal_L34e	8.00E-40
55	2	gi|7291886| RpL19 gene product	2.00E-79	gi|7291886	pfam01280 Ribosomal_L19e	7.00E-70
57	2	gi|7291730| CG3195 gene product	5.00E-68	gi|7291730	pfam00298 Ribosomal_L11	6.00E-38
92	1	gi|7293042| CG9091 gene product	1.00E-35	gi|7293042	pfam01907 Ribosomal_L37e	6.00E-22
102	1	gi|7300661| RpS20 gene product	3.00E-51	gi|7300661	pfam00338 Ribosomal_S10	4.00E-33
111	1	gi|10439989| unnamed protein product...	4.00E-23	gi|10439989	No matches found	
126	1	gi|7291436| CG4046 gene product	2.00E-65	gi|7291436	pfam00380 Ribosomal_S9	1.00E-41
127	1	gi|7289746| Qm gene product [alt 1]	2.00E-72	gi|7289746	pfam00826 Ribosomal_L10e	2.00E-66
131	1	gi|7291277| RpL29 gene product	3.00E-22	gi|7291277	pfam01779 Ribosomal_L29e	5.00E-13
132	1	gi|1359478| put. S3a ribosomal protein	4.00E-76	gi|1359478	pfam01015 Ribosomal_S3Ae	9.00E-47
142	1	gi|7290749| CG3203 gene product [alt	8.00E-66	gi|7290749	pfam00237 Ribosomal_L22	4.00E-10
145	1	gi|5690418| ribosomal protei...	2.00E-44	gi|5690418	pfam00833 Ribosomal_S17e	8.00E-32
160	1	gi|7290065| RpL36 gene product	4.00E-32	gi|7290065	pfam01158 Ribosomal_L36e	7.00E-35
166	1	gi|7292864| CG2033 gene product [alt	3.00E-65	gi|7292864	pfam00410 Ribosomal_S8	6.00E-37
170	1	gi|9438110| putative large s...	1.00E-45	gi|9438110	pfam00935 Ribosomal_L44	9.00E-24
171	1	gi|7301850| CG7808 gene product	3.00E-64	gi|7301850	pfam01201 Ribosomal_S8e	9.00E-52
191	1	gi|7301368| CG4759 gene product	8.00E-37	gi|7301368	pfam01777 Ribosomal_L27e	3.00E-40
222	1	gi|14585747| ribosomal protein L39	3.00E-19	gi|14585747	pfam00832 Ribosomal_L39	2.00E-15
250	1	gi|1245447| putative ribosomal protein	1.00E-12	gi|1245447	No matches found	
251	1	gi|7293850| CG6846 gene product	1.00E-46	gi|7293850	pfam00467 Ribosomal_L24	4.00E-14
207	1	gi|7291445| CG3751 gene product	4.00E-52	gi|7291445	pfam01282 Ribosomal_S24e	7.00E-29
172	1	gi|7295253| Srp19 gene product	3.00E-17	gi|7295253	pfam01922 SRP19	2.00E-12
230	1	gi|7295014| CG5651 gene product	2.00E-95	gi|7295014	pfam00005 ABC_tran	1.00E-24
217	1	gi|4389443| SPC 21-kDa-like	4.00E-63	gi|4389443	pfam00461 Peptidase_S26	2.00E-30
236	1	gi|12835580| putative [Mus musculus]	4.00E-08	gi|12835580	No matches found	
243	1	gi|7290606| CG3187 gene product	8.00E-69	gi|7290606	LOAD_sir2 sir2	7.00E-32
78	1	gi|7299412| CG6666 gene product	3.00E-34	gi|7299412	pfam01127 Sdh_cyt	2.00E-33
167	1	gi|7303341| CG4670 gene product	7.00E-22	gi|7303341	pfam01500 Keratin_B2	0.002
31	3	gi|7716428| thioredoxin 1 [A...	7.00E-11	gi|7716428	pfam00085 thiored	0.002
247	1	gi|6469517| translation init...	1.00E-45	gi|6469517	pfam01287 eIF-5a	2.00E-17
47	2	gi|7299413| CG4800 gene product	3.00E-57	gi|7299413	pfam00838 TCTP	4.00E-55
225	1	gi|7290103| EG:BACR7A4.5 gene product	2.00E-27	gi|7290103	pfam00033 cytochrome_b_N	0.009
165	1	gi|7300467| CG14285 gene product	6.00E-07	gi|7300467	No matches found	
209	1	gi|10728827| VhaSFD gene product [alt	1.00E-61	gi|10728827	pfam01602 Adaptin_N	1.00E-05
229	1	gi|1220128| vacuolar ATPase [Anopheles	8.00E-57	gi|1220128	pfam01990 ATP-synt_F	6.00E-31
244	1	gi|10728827| VhaSFD gene product [alt	2.00E-18	gi|10728827	pfam00607 gag_p24	0.004
125	1	gi|7303921| CG8029 gene product [alt	2.00E-08	gi|7303921	No matches found	
240	1	gi|9502404| Hypothetical zin...	7.00E-30	gi|9502404	pfam00096 zf-C2H2	1.00E-05

[1]Clus, cluster number.
[2]R, number of sequences of a given custer (R, representation).
[3]GenBank match, best match to the GenBank database.
[4]E-value, indicates significance of match to NR sequence of previous column.
[5]Gi numb, GenBank accession number.
[6]PFAM match, best match to the PFAM database.
[7]E-value, indicates significance of match to CDD sequence of previous column.

Results and Discussion

A literature search indicates that the salivary gland of *Anopheles gambiae* contains a number of putative proteins including apyrase, 5′ nucleotidase, lyzozyme, members of the D7 family of proteins, nine so-called *A. gambiae* hypothetical proteins (HP), and eight so-called *A. gambiae* salivary gland (SG) proteins (Arcà et al., 1999). Most of these proteins are translations of expressed sequence tags (EST) or full-length clones; accordingly, they have not been unambiguously characterized as salivary molecules. In an

Table 1. *Continued*

Org[8]	N[9]	Function[10]	F[11]	Clone[12]	FASTA[13]	CLUSTAL[14]
Rattus norvegicus	Y	Ribosomal protein	H	NoORF	AGN10E60-52clu.txt	AGN10E60-52aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NoORF	AGN10E60-53clu.txt	AGN10E60-53aln.txt
Aedes triseriatus	Y	Ribosomal protein	H	NoORF	AGN10E60-54clu.txt	AGN10E60-54aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NoORF	AGN10E60-55clu.txt	AGN10E60-55aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-57clu.txt	AGN10E60-57aln.txt
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-92clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NoORF	AGN10E60-102clu.txt	
Homo sapiens	Y	Ribosomal protein	H	NoORF	AGN10E60-111clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-126clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-127clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-131clu.txt	
Anopheles gambiae	N	Ribosomal protein	H	NOSIG	AGN10E60-132clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-142clu.txt	
Anopheles gambiae	N	Ribosomal protein	H	NOSIG	AGN10E60-145clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NoORF	AGN10E60-160clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-166clu.txt	
Aedes triseriatus	Y	Ribosomal protein	H	NOSIG	AGN10E60-170clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-171clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-191clu.txt	
Culex pipiens pallens	Y	Ribosomal protein	H	NoORF	AGN10E60-222clu.txt	
Anopheles gambiae	N	Ribosomal protein	H	NoORF	AGN10E60-250clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-251clu.txt	
Drosophila melanogaster	Y	Ribosomal protein	H	NOSIG	AGN10E60-207clu.txt	
Drosophila melanogaster	Y	Ribosomal signal recognition	H	NoORF	AGN10E60-172clu.txt	
Drosophila melanogaster	Y	RNAse L inhibitor	H	NOSIG	AGN10E60-230clu.txt	
Drosophila melanogaster	Y	Signal peptidase	H	ANCH	AGN10E60-217clu.txt	
Mus musculus	Y	Similar to yeast Maf1p	H	NOSIG	AGN10E60-236clu.txt	
Drosophila melanogaster	Y	Sir2 family	H	NOSIG	AGN10E60-243clu.txt	
Drosophila melanogaster	Y	Succinate dehydrogenase cytochrome b subunit	H	NoORF	AGN10E60-78clu.txt	
Drosophila melanogaster	Y	Sulfhydryl oxidase	H	NoORF	AGN10E60-167clu.txt	
Anopheles gambiae	N	Thioredoxin	H	NoORF	AGN10E60-31clu.txt	AGN10E60-31aln.txt
Spodoptera frugiperda	Y	Translation initiation factor	H	NOSIG	AGN10E60-247clu.txt	
Drosophila melanogaster	Y	Translationally controlled tumour protein	H	NOSIG	AGN10E60-47clu.txt	AGN10E60-47aln.txt
Drosophila melanogaster	Y	Translocation protein	H	NOSIG	AGN10E60-225clu.txt	
Drosophila melanogaster	Y	Unknown protein – conserved	H	NOSIG	AGN10E60-165clu.txt	
Drosophila melanogaster	Y	V-ATPase subunit	H	NoORF	AGN10E60-209clu.txt	
Anopheles gambiae	N	V-ATPase subunit	H	NOSIG	AGN10E60-229clu.txt	
Drosophila melanogaster	Y	V-ATPase subunit	H	NOSIG	AGN10E60-244clu.txt	
Drosophila melanogaster	Y	V-ATPase synthase	H	NoORF	AGN10E60-125clu.txt	
Homo sapiens	Y	Zinc finger protein	H	NoORF	AGN10E60-240clu.txt	

[8]Org, organism where a match was found.
[9]N, Novel (Y, yes; N, no).
[10]Function, putative function or biological property.
[11]F, Function; H, housekeeping.
[12]Clone: NOSIG, no signal peptide detected by SignalP server; No ORF, absence of a open-reading frame; ANCH, anchor protein.
[13]FASTA, FASTA-formatted sequences.
[14]CLUSTAL, clustal alignment for two or more sequences.

attempt to improve our understanding of the complexity of the proteins and transcripts expressed in *A. gambiae* salivary glands, we have performed SDS-PAGE and a cDNA library using, respectively, the proteins and mRNA from this same tissue.

SDS-PAGE of A. gambiae *salivary proteins*

Fig. 1 shows the pattern of separation of *A. gambiae* salivary protein by SDS-PAGE stained by Coomassie Blue. The gel shows relatively low tissue complexity, with approximately 15 clearly visible stained bands and many others lightly stained.

Table 2. Anopheles gambiae *secretory cDNAs*

Clus[1]	R[2]	GenBank match[3]	E-value[4]	Gi numb[5]	PFAM match[6]	E-value[7]
6	33	gi|2114497| 30 kDa salivary gland	7.00E-07	gi|2114497	pfam01127 Sdh_cyt	7.00E-07
58	2	gi|159550| amylase [Aedes aegypti]	1.00E-66	gi|159550	Smart smart00632 Aamy_C	1.00E-14
105	1	gi|2285788| Multiprotein bridging fac...	5.00E-51	gi|2285788	pfam01381 HTH_3	3.00E-06
17	6	gi|4127344| cE5 protein [Anopheles	7.00E-20	gi|4127344	pfam01028 Topoisomerase_I	1.00E-09
14	8	gi|8927462| antigen 5 precur...	3.00E-36	gi|8927462	Smart smart00198 SCP	9.00E-30
69	1	gi|7292977| CG9400 gene product	1.00E-18	gi|7292977	Smart smart00198 SCP	2.00E-20
210	1	gi|4887102| antigen 5-relate...	3.00E-08	gi|4887102	Smart smart00198 SCP	9.00E-08
183	1	gi|7299219| Crc gene product	5.00E-87	gi|7299219	pfam00262 calreticulin	3.00E-70
15	7	gi|159559| D7 protein [Aedes aegypti]	4.00E-23	gi|159559	No matches found	
8	23	gi|4538887| D7-related 1 protein [Ano...	9.00E-82	gi|4538887	pfam01604 7tm_5	9.00E-05
208	1	gi|4538887| D7-related 1 protein [Ano...	7.00E-07	gi|4538887	No matches found	
88	1	gi|4538889| D7-related 2 protein [Ano...	4.00E-05	gi|4538889	No matches found	
1	49	gi|4538889| D7-related 2 protein [Ano...	8.00E-77	gi|4538889	pfam02326 YMF19	0.001
2	47	gi|4538891| D7-related 3 protein [Ano...	4.00E-94	gi|4538891	pfam01028 Topoisomerase_I	3.00E-05
11	12	gi|13537670| D7r4 protein [Anopheles...	3.00E-94	gi|13537670	pfam01598 Sterol_desat	0.007
26	3	gi|13537670| D7r4 protein [Anopheles...	4.00E-17	gi|13537670	No matches found	
56	2	gi|13537664| gSG1-like 2 protein [Ano...	1.00E-126	gi|13537664	pfam01500 Keratin_B2	5.00E-04
5	33	gi|13537660| gSG2-like protein [Anoph...	3.00E-30	gi|13537660	No matches found	
13	9	gi|13537662| gSG5 protein [Anopheles...	6.00E-93	gi|13537662	Smart smart00052 DUF2	0.005
9	20	gi|13537666| gSG6 protein [Anopheles...	3.00E-55	gi|13537666	pfam01688 Herpes_gI	0.002
12	10	gi|13537668| gSG7 protein [Anopheles...	3.00E-77	gi|13537668	No matches found	
19	6	gi|13537668| gSG7 protein [Anopheles...	3.00E-09	gi|13537668	No matches found	
22	5	gi|13537672| hypothetical protein [An...	2.00E-74	gi|13537672	pfam00710 Asparaginase	6.00E-04
7	31	gi|4127307| hypothetical protein	7.00E-05	gi|4127307	pfam01028 Topoisomerase_I	0.002
10	13	gi|894206| lysozyme [Anopheles gambiae]	2.00E-71	gi|894206	Smart smart00263 LYZ1	1.00E-52
34	2	No matches found	pfam01456 Tryp_mucin		2.00E-07	
16	6	gi|4582528| putative 5′-nucleotidase...	1.00E-129	gi|4582528	pfam02872 5_nucleotidaseC	1.00E-49
46	2	gi|4582528| putative 5′-nucleotidase...	3.00E-73	gi|4582528	pfam01009 5_nucleotidase	1.00E-24
20	6	gi|4582526| putative apyrase [Anophel...	2.00E-41	gi|4582526	pfam02872 5_nucleotidaseC	2.00E-06
128	1	gi|7293955| CG7484 gene product	2.00E-31	gi|7293955	No matches found	
212	1	gi|4210615| SG1 protein [Anopheles ga...	4.00E-07	gi|4210615	No matches found	
37	2	gi|4210615| SG1 protein [Anopheles ga...	2.00E-10	gi|4210615	pfam02122 Luteo_ORF2	3.00E-04
21	5	gi|4210617| SG2 protein [Anopheles ga...	1.00E-23	gi|4210617	pfam01028 Topoisomerase_I	2.00E-09
4	38	No matches found			No matches found	
27	3	No matches found			No matches found	
68	1	No matches found			No matches found	
83	1	No matches found			No matches found	
90	1	No matches found			pfam02326 YMF19	0.005
93	1	No matches found			No matches found	
134	1	No matches found			No matches found	
139	1	No matches found			pfam01604 7tm_5	0.007
157	1	No matches found			No matches found	
162	1	No matches found			No matches found	
188	1	No matches found			No matches found	
29	3	No matches found			pfam01490 Aa_trans	0.002

[1]Clus, cluster number.
[2]R, number of sequences of a given custer (R, representation).
[3]GenBank match, best match to the GenBank database.
[4]E-value, indicates significance of match to NR sequence of previous column.
[5]Gi numb, GenBank accession number.
[6]PFAM match, best match to the PFAM database.
[7]E-value, indicates significance of match to CDD sequence of previous column.

Table 2. *Continued*

Org[8]	N[9]	Function[10]	F[11]	Clone[12]	FASTA[13]	CLUSTAL[14]
Aedes aegypti	Y	30 kDa allergen family	S	SIG	AGN10E60-6clu.txt	AGN10E60-6aln.txt
Aedes aegypti	Y	Amylase	S	NoORF	AGN10E60-58clu.txt	AGN10E60-58aln.txt
Bombyx mori	Y	Angiogenesis function? –Similar to human EDF-1	S	NoORF	AGN10E60-105clu.txt	
Anopheles gambiae	N	Anticlotting	S	SIG	AGN10E60-17clu.txt	AGN10E60-17aln.txt
Glossina morsitans morsitans	Y	Antigen 5 family	S	SIG	AGN10E60-14clu.txt	AGN10E60-14aln.txt
Drosophila melanogaster	Y	Antigen 5 protein	S	SIG	AGN10E60-69clu.txt	
Lutzomyia longipalpis	Y	Antigen 5 protein	S	NOSIG	AGN10E60-210clu.txt	
Drosophila melanogaster	Y	Calreticulin	S	SIG	AGN10E60-183clu.txt	
Aedes aegypti	Y	D7 protein	S	SIG	AGN10E60-15clu.txt	AGN10E60-15aln.txt
Anopheles gambiae	N	D7r1 protein	S	SIG	AGN10E60-8clu.txt	AGN10E60-8aln.txt
Anopheles gambiae	Y	D7r1 protein new member	S	NoORF	AGN10E60-208clu.txt	
Anopheles gambiae	Y	D7r2 family – New member	S	NoORF	AGN10E60-88clu.txt	
Anopheles gambiae	N	D7r2 protein	S	SIG	AGN10E60-1clu.txt	AGN10E60-1aln.txt
Anopheles gambiae	N	D7r3 protein	S	SIG	AGN10E60-2clu.txt	AGN10E60-2aln.txt
Anopheles gambiae	N	D7r4 protein	S	SIG	AGN10E60-11clu.txt	AGN10E60-11aln.txt
Anopheles gambiae	Y	D7r protein new member	S	SIG	AGN10E60-26clu.txt	AGN10E60-26aln.txt
Anopheles gambiae	N	gSG1-like 2 protein	S	NoORF	AGN10E60-56clu.txt	AGN10E60-56aln.txt
Anopheles gambiae	N	gSG2 protein	S	SIG	AGN10E60-5clu.txt	AGN10E60-5aln.txt
Anopheles gambiae	N	gSG5 protein	S	SIG	AGN10E60-13clu.txt	AGN10E60-13aln.txt
Anopheles gambiae	N	gSG6 protein	S	SIG	AGN10E60-9clu.txt	AGN10E60-9aln.txt
Anopheles gambiae	N	gSG7 protein	S	SIG	AGN10E60-12clu.txt	AGN10E60-12aln.txt
Anopheles gambiae	Y	gSG7 protein – new member?	S	SIG	AGN10E60-19clu.txt	AGN10E60-19aln.txt
Anopheles gambiae	N	HP gi|13537672	S	ANCH	AGN10E60-22clu.txt	AGN10E60-22aln.txt
Anopheles gambiae	N	HP gi|4127307	S	SIG	AGN10E60-7clu.txt	AGN10E60-7aln.txt
Anopheles gambiae	N	Lyzozyme	S	SIG	AGN10E60-10clu.txt	AGN10E60-10aln.txt
	Y	Mucin	S	SIG	AGN10E60-34clu.txt	AGN10E60-34aln.txt
Anopheles gambiae	N	Putative 5′-nuc truncated clone	S	NoORF	AGN10E60-16clu.txt	AGN10E60-16aln.txt
Anopheles gambiae	N	Putative 5′-nucleotidase	S	SIG	AGN10E60-46clu.txt	AGN10E60-46aln.txt
Anopheles gambiae	N	Putative apyrase – truncated clone	S	NOSIG	AGN10E60-20clu.txt	AGN10E60-20aln.txt
Drosophila melanogaster	Y	Selenoprotein precursor	S	SIG	AGN10E60-128clu.txt	
Anopheles gambiae	Y	SG1 protein – New member?	S	NOSIG	AGN10E60-212clu.txt	
Anopheles gambiae	Y	SG1 protein – New member?	S	NOSIG	AGN10E60-37clu.txt	AGN10E60-37aln.txt
Anopheles gambiae	N	SG2	S	SIG	AGN10E60-21clu.txt	AGN10E60-21aln.txt
	Y	Unknown protein	S	SIG	AGN10E60-4clu.txt	AGN10E60-4aln.txt
	Y	Unknown protein	S	SIG	AGN10E60-27clu.txt	AGN10E60-27aln.txt
	Y	Unknown protein	S	SIG	AGN10E60-68clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-83clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-90clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-93clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-134clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-139clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-157clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-162clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-188clu.txt	
	Y	Unknown protein	S	SIG	AGN10E60-29clu.txt	AGN10E60-29aln.txt

[8]Org, organism where a match was found.
[9]N, Novel (Y, yes; N, no).
[10]Function, putative function or biological property.
[11]F, Function; S, secretory
[12]Clone: NOSIG, no signal peptide detected by SignalP server; No ORF, absence of a open-reading frame; ANCH, anchor protein.
[13]FASTA, FASTA-formatted sequences.
[14]CLUSTAL, clustal alignment for two or more sequences.

To identify these proteins, they were transferred to PVDF membranes and the bands cut from the membrane and submitted to Edman degradation. Amino-terminal information was successfully obtained for many of these bands, and they were identified as SG6 (approx. 10 kDa apparent molecular mass), *A. gambiae* D7-related proteins 1-3 (approx. 10–14 kDa apparent molecular mass), similar to *Glossinia morsitans* antigen 5 (approx. 30 kDa), similar to *Aedes aegypti* D7 (approx. 33 kDa), similar to *A. aegypti* 30 kDa allergen (approx. 36 kDa), HP 8 (CB1, 44 kDa), similar to HP 9 (bB2, 46 kDa) and herein called HP 9-like, SG1-like 2 (approx. 48 kDa), putative 5′ nucleotidase (approx. 64 kDa) and SG1 (approx. 105 kDa). Although the predicted translation products of some of these proteins have been reported (Arcà et al., 1999), amino acid sequencing has not been performed before. Edman degradation for other bands was attempted unsuccessfully, either because the protein's amino terminus was blocked, or because PTH-amino acids could not be reliably identified.

cDNA library of the salivary gland of A. gambiae

To complement the data generated by SDS-PAGE and identify potentially novel molecules in the salivary gland of *A. gambiae*, a cDNA library was constructed and hundreds of independent clones randomly 5′ sequenced. When a cluster analysis of all 691 sequences from this library was performed at e^{-60}, 251 independent clusters were organized. Subsequently, clusters were blasted against the nonredundant (NR) and protein motifs databases. Signal peptides were predicted by submission of the cluster sequences to the SignalP server, allowing the identification of putative secretory (S) and housekeeping (H) cDNA. A comprehensive diagram depicting the steps used for generation of the data is shown in Fig. 2. The results are presented as Tables 1–4. The electronic versions of the tables, available on request, also contain: (i) columns with hyperlinks to the best match of the NR database, (ii) links to NR matches found for the cluster, (iii) matches to the conserved domain database (CDD), (iv) FASTA-formatted files for each cluster, and (v) CLUSTAL alignments of each cluster having two or more sequences.

Fig. 3A shows that of the 691 sequences are concerned, 403 (58.3 %) code for putative S proteins, 165 (23.8 %) for housekeeping proteins and 123 (17.8 %) for proteins that could not be identified as housekeeping or secretory (unknown, U). Accordingly, cDNA for secretory proteins are highly represented in our library, suggesting that *in vivo* these molecules are preferentially expressed over H and U proteins. Fig. 3B shows that of the 251 clusters (including H, S and U), 127 (40.6 %) match sequences related to *Drosophila melanogaster* or other organisms; however, only 41 (16.3 %) have been assigned exclusively to the *A. gambiae* salivary gland. This indicates that 120 clusters (49.4 %) lack NCBI hits, although it is possible that related nucleotide sequences have been deposited as EST in other databases.

cDNAs coding for putative housekeeping proteins

Table 1 describes *A. gambiae* cDNA sequences with probable housekeeping function found in our database. These include many different ribosomal proteins, t-RNA synthases, cytochrome oxidase, elongation and translation factors, endoplasmic reticulum proteins, NADH dehydrogenases, heat-shock protein, actin depolarization factor, arrestin, aminotransferase, clatrin and porin gene product. Many enzymes linked to respiratory metabolism or mitochondria proteins were identified, including adenosine diphosphatase, ADP/ATP carrier protein, unnamed protein product, V-ATPase and Na^+K^+ ATPase subunits. Other enzymes or proteins were nucleoside diphosphate kinase, ornithine decarboxylase, peptide chain release factor, prenylated rab acceptor, RNAse L inhibitor proteasome protein, KE2 family, CG14525, Mago Nashi protein, β-glucosidase and condensation domain. Of interest, clusters possessing matches to tetraspanin, hemopexin, heat-shock protein, TRIO protein, multiprotein bridging factor (MBF) and antioxidant molecules have also been found; their putative function is discussed below.

Tetraspanins, transmembrane proteins first discovered on the surface of human leukocytes, have previously been identified in *Drosophila melanogaster*, *Caenorhabditis elegans*, *Apis mellifera* and *Manduca sexta*. This is the first description of tetraspanin in a mosquito. Their function is not precisely known, but data from biochemical studies and knockout mice suggest that they play a major role in membrane biology, operating as molecular facilitators of diverse cellular functions from cell adhesion to signal transduction (Todres et al., 2000). Of interest, the tetraspanin CD9 associates with the CD36, the *Plasmodium falciparum* receptor on platelets and endothelial cells (Miao et al., 2001). Whether salivary gland tetraspanin has any role in cell–parasite interactions remains to be determined.

Also noteworthy is the identification of clones coding for proteins with antioxidant function. This paper reports the first identification of hemopexin (hpx) in the salivary gland of a blood-sucking insect. Hpx is a haem-binding plasma glycoprotein that forms a line of defense against hemoglobin-mediated oxidative damage during hemolysis (Delanghe and Langlois, 2001). In fact, hpx complexes with heme noncovalently with high affinity (K_d<1 pmol l^{-1}) and shows much lower peroxidase- and catalyse-like activity than the nonprotein heme. In addition, hpx heme binds nitric oxide (NO) and carbon monoxide (CO) and may protect against NO-mediated toxicity, especially in conditions of hemolysis. Hpx is thus a molecule that safely carries heme. Perhaps it is present in the salivary gland due to the synthesis of relatively large amounts of salivary peroxidase, which function as a vasodilator. In addition, a clone was identified coding for thioredoxin (Thrx), a molecule that plays a fundamental role in maintaining a reducing cellular milieu together with Thrx reductase (ThrxR) and NADPH (Holmgren and Bjornstedt, 1995). Interestingly, ThrxR has both Thrx and protein disulphide isomerase (PDI) as substrate (Nakamura et al., 1997), and a clone for PDI is in our library. We conclude that components of the Thrx system are present in the salivary

gland of *A. gambiae* and that they may operate in concert with Hpx and other molecules to prevent haem-driven free radical attack, considering that this organ is actively engaged in hemeprotein synthesis. Finally, we have found clones coding for heat-shock proteins, a family of proteins that functions as chaperones, or are involved in cell defense against external stressors from various sources (Lund, 2001). In fact, a general function of heat-shock proteins is to prevent protein misfolding and aggregation in highly crowded cellular environments or under conditions of denaturing stress (Young et al., 2001).

We have also found clones with sequence homology to signaling molecules. TRIO is a multidomain protein that binds the lymphocyte activating receptor transmembrane tyrosine phosphatase (PTPase) and contains a protein kinase domain. It has been proposed that TRIO may orchestrate cell-matrix and cytoskeletal rearrangements necessary for cell migration (Lin and Greenberg, 2000). Although we have found a signal peptide for *A. gambiae* TRIO, our alignment with *Drosophila* TRIO (a protein of approx. 200 kDa with no secretion sequence) makes it uncertain whether this mosquito form of TRIO is, in fact, a secreted protein or a truncated protein with a false-positive signal peptide. Finally, an open-reading frame with the complete coding region for a protein with homology to *Bombyx mori* MBF without signal peptide has been identified (Takemaru et al., 1997). MBF is similar to endothelial cell differentiation factor (Dragoni et al., 1998), an intracellular protein that plays a role in regulation of human endothelial cell functions including formation of blood vessels. The precise functions of TRIO and MBP in the salivary gland of *A. gambiae* remain to be determined.

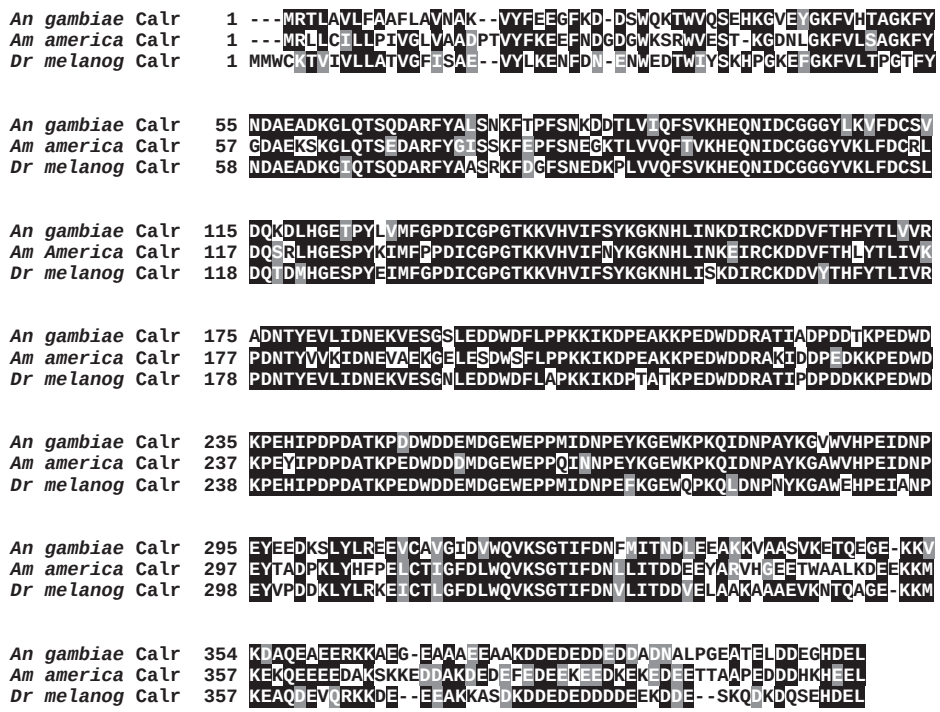

Fig. 4. CLUSTAL alignment of novel *A. gambiae* calreticulin (*An gambiae* Calr, gi 18389888), *Amblyomma americanum* calreticulin (*Am america* Calr, gi 3924593), and *D. melanogaster* calreticulin (*Dr melanog* Calr, gi 7299219). Similar amino acid residues are marked with a gray background, identical amino acids with a black background.

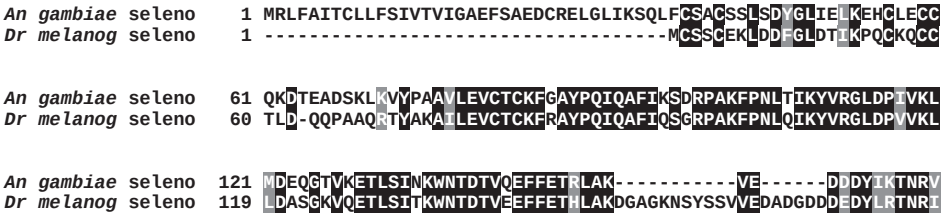

Fig. 5. CLUSTAL alignment of novel *A. gambiae* selenoprotein (*An gambiae* seleno; gi 18389880) and *D. melanogaster* (*Dr melanog* seleno, gi 7293955). Similar amino acid residues are marked with a gray background, identical amino acids with a black background.

cDNAs coding for putative secretory proteins

Table 2 shows clusters probably associated with secreted products. Some match sequences have already been reported for the *A. gambiae* salivary gland; however, several are novel with database hits to genes unrelated to *A. gambiae*, or without database hits.

A cDNA has been identified having an open reading frame with signal peptide and sequence homology to calreticulin, an ubiquitous intracellular protein present in the sarcoplasmic reticulum and involved in calcium homeostasis (Johnson et al., 2001). Calreticulin has also been identified extracellularly in the supernatant of Epstein-Barr virus-immortalized cells; this secreted form has been shown to inhibit angiogenesis, the biological process by which new blood vessels are formed (Pike et al., 1998). This suggests that the saliva of *A. gambiae* may inhibit endothelial cell proliferation, a proinflammatory event associated with host response to injury, and other proinflammatory responses (Griffioen and Molema, 2000). The CLUSTAL alignment of calreticulin from *A. gambiae*, *Amblyomma americanum* and *D. melanogaster* is shown in Fig. 4.

Another full-length clone containing a typical secretion sequence and homologous to selenoproteins has been identified. The selenoproteins incorporate selenocysteine, a

Table 3. Anopheles gambiae *unknown function cDNAs*

Clus[1]	R[2]	GenBank match[3]	E-value[4]	Gi numb[5]	PFAM match[6]	E-value[7]
72	1	No matches found			pfam02632 BioY	0.003
130	1	No matches found			No matches found	
140	1	No matches found			No matches found	
155	1	No matches found			No matches found	
203	1	gi|7300136| ARP-like gene product	9.00E-52	gi|7300136	No matches found	
119	1	gi|13537674| gSG8 protein [Anopheles...	7.00E-58	gi|13537674	No matches found	
198	1	gi|7300660| CG17271 gene product	5.00E-07	gi|7300660	No matches found	
187	1	gi|7299765| CG9796 gene product	5.00E-39	gi|7299765	No matches found	
221	1	gi|12851155| putative [Mus musculus]	5.00E-25	gi|12851155	pfam00361 oxidored_q1	8.00E-04
25	4	No matches found			No matches found	
32	3	No matches found			pfam02326 YMF19	0.005
35	2	No matches found			No matches found	
41	2	No matches found			No matches found	
44	2	No matches found			No matches found	
59	2	No matches found			No matches found	
62	1	No matches found			No matches found	
63	1	No matches found			No matches found	
64	1	No matches found			pfam02695 DUF216	1.00E-07
66	1	No matches found			No matches found	
67	1	No matches found			No matches found	
71	1	No matches found			No matches found	
75	1	No matches found			pfam00001 7tm_1	0.006
76	1	No matches found			No matches found	
77	1	No matches found			No matches found	
80	1	No matches found			No matches found	
81	1	No matches found			No matches found	
82	1	No matches found			pfam02361 CbiQ	0.009
85	1	No matches found			No matches found	
86	1	No matches found			No matches found	
89	1	No matches found			No matches found	
91	1	No matches found			No matches found	
94	1	No matches found			No matches found	
95	1	No matches found			No matches found	
96	1	No matches found			No matches found	
97	1	No matches found			No matches found	
98	1	No matches found			No matches found	
100	1	No matches found			No matches found	
101	1	No matches found			No matches found	
107	1	No matches found			No matches found	
108	1	No matches found			No matches found	
110	1	No matches found			No matches found	
113	1	No matches found			No matches found	
114	1	No matches found			pfam00902 UPF0032	0.002
115	1	No matches found			No matches found	
117	1	No matches found			No matches found	
118	1	No matches found			No matches found	
122	1	No matches found			pfam01306 LacY_symp	0.005
124	1	No matches found			pfam01028 Topoisomerase_I	9.00E-04
129	1	No matches found			No matches found	
133	1	No matches found			No matches found	
135	1	No matches found			No matches found	
136	1	No matches found			No matches found	
137	1	No matches found			pfam01604 7tm_5	0.004
138	1	gi|6467825| Spen RNP motif p...	0.004	gi|6467825	No matches found	
141	1	No matches found			No matches found	

Continued on p. 2444.

Table 3. *Continued*

Org[8]	N[9]	Function[10]	F[11]	Clone[12]	FASTA[13]	CLUSTAL[14]
	Y	Unknown protein	U	ANCH	AGN10E60-72clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-130clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-140clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-155clu.txt	
Drosophila melanogaster	Y	Arginine rich protein	U	NoORF	AGN10E60-203clu.txt	
Anopheles gambiae	N	gSG8 protein	U	NoORF	AGN10E60-119clu.txt	
Drosophila melanogaster	Y	Similar to adenyl-cyclase and sex determining protein	U	NoORF	AGN10E60-198clu.txt	
Drosophila melanogaster	Y	Similar to Asn-specific plant protease	U	NoORF	AGN10E60-187clu.txt	
Mus musculus	Y	Ubiquinone complex protein	U	ANCH	AGN10E60-221clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-25clu.txt	AGN10E60-25aln.txt
	Y	Unknown protein	U	NoORF	AGN10E60-32clu.txt	AGN10E60-32aln.txt
	Y	Unknown protein	U	NoORF	AGN10E60-35clu.txt	AGN10E60-35aln.txt
	Y	Unknown protein	U	NoORF	AGN10E60-41clu.txt	AGN10E60-41aln.txt
	Y	Unknown protein	U	NoORF	AGN10E60-44clu.txt	AGN10E60-44aln.txt
	Y	Unknown protein	U	NoORF	AGN10E60-59clu.txt	AGN10E60-59aln.txt
	Y	Unknown protein	U	NoORF	AGN10E60-62clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-63clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-64clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-66clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-67clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-71clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-75clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-76clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-77clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-80clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-81clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-82clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-85clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-86clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-89clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-91clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-94clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-95clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-96clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-97clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-98clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-100clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-101clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-107clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-108clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-110clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-113clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-114clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-115clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-117clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-118clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-122clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-124clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-129clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-133clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-135clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-136clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-137clu.txt	
Drosophila melanogaster	Y	Unknown protein	U	NOSIG	AGN10E60-138clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-141clu.txt	

Table 3. *Continued*

Clus[1]	R[2]	GenBank match[3]	E-value[4]	Gi numb[5]	PFAM match[6]	E-value[7]
143	1	No matches found			No matches found	
144	1	No matches found			No matches found	
148	1	gi|7290682| CG3599 gene product	0.009	gi|7290682	No matches found	
149	1	gi|7303810| CG2249 gene product	8.00E-07	gi|7303810	No matches found	
150	1	No matches found			No matches found	
151	1	No matches found			No matches found	
152	1	No matches found			No matches found	
153	1	gi|6560685| unknown [Manduca...	2.00E-13	gi|6560685	No matches found	
154	1	No matches found			No matches found	
156	1	No matches found			No matches found	
158	1	No matches found			No matches found	
159	1	No matches found			No matches found	
161	1	No matches found			No matches found	
173	1	No matches found			No matches found	
174	1	gi|7303552| CG13189 gene product	9.00E-21	gi|7303552	No matches found	
175	1	gi|7296046| CG17652 gene product	0.008	gi|7296046	No matches found	
176	1	No matches found			No matches found	
177	1	No matches found			No matches found	
178	1	No matches found			No matches found	
180	1	No matches found			No matches found	
185	1	No matches found			No matches found	
186	1	gi|7303876| CG12929 gene product	8.00E-22	gi|7303876	pfam00892 DUF6	0.002
189	1	No matches found			No matches found	
190	1	No matches found			No matches found	
192	1	No matches found			No matches found	
193	1	No matches found			No matches found	
194	1	No matches found			No matches found	
196	1	No matches found			No matches found	
197	1	No matches found			No matches found	
200	1	No matches found			No matches found	
202	1	gi|309071| ribosomal protein S7	0.065	gi|309071	No matches found	
204	1	No matches found			No matches found	
205	1	No matches found			No matches found	
206	1	No matches found			No matches found	
211	1	No matches found			No matches found	
213	1	No matches found			pfam01813 ATP-synt_D	0.003
214	1	No matches found			No matches found	
215	1	No matches found			pfam00001 7tm_1	7.00E-04
218	1	No matches found			No matches found	
219	1	No matches found			pfam01604 7tm_5	0.007
220	1	No matches found			No matches found	
224	1	No matches found			pfam01028 Topoisomerase_I	0.003
227	1	No matches found			No matches found	
231	1	No matches found			No matches found	
233	1	No matches found			No matches found	
234	1	No matches found			No matches found	
237	1	No matches found			pfam00902 UPF0032	2.00E-04
238	1	No matches found			No matches found	
241	1	No matches found			No matches found	
242	1	No matches found			No matches found	
245	1	No matches found			No matches found	
246	1	No matches found			No matches found	
248	1	gi|13421488| 2-oxoglutarate	0.002	gi|13421488	No matches found	
163	1	gi|7303010| CG8386 gene product	5.00E-56	gi|7303010	No matches found	
184	1	gi|7295175| CG8209 gene product	4.00E-24	gi|7295175	pfam00627 UBA	2.00E-08

Continued on p. 2446.

Table 3. *Continued*

Org[8]	N[9]	Function[10]	F[11]	Clone[12]	FASTA[13]	CLUSTAL[14]
	Y	Unknown protein	U	NoORF	AGN10E60-143clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-144clu.txt	
Drosophila melanogaster	Y	Unknown protein	U	NOSIG	AGN10E60-148clu.txt	
Drosophila melanogaster	Y	Unknown protein	U	NOSIG	AGN10E60-149clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-150clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-151clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-152clu.txt	
Manduca sexta	Y	Unknown protein	U	NoORF	AGN10E60-153clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-154clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-156clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-158clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-159clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-161clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-173clu.txt	
Drosophila melanogaster	Y	Unknown protein	U	NoORF	AGN10E60-174clu.txt	
Drosophila melanogaster	Y	Unknown protein	U	NoORF	AGN10E60-175clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-176clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-177clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-178clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-180clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-185clu.txt	
Drosophila melanogaster	Y	Unknown protein	U	NoORF	AGN10E60-186clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-189clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-190clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-192clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-193clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-194clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-196clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-197clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-200clu.txt	
Anopheles gambiae	N	Unknown protein	U	NoORF	AGN10E60-202clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-204clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-205clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-206clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-211clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-213clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-214clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-215clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-218clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-219clu.txt	
	Y	Unknown protein	U	ANCH	AGN10E60-220clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-224clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-227clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-231clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-233clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-234clu.txt	
	Y	Unknown protein	U	NOSIG	AGN10E60-237clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-238clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-241clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-242clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-245clu.txt	
	Y	Unknown protein	U	NoORF	AGN10E60-246clu.txt	
Caulobacter crescentus	Y	Unknown protein	U	NoORF	AGN10E60-248clu.txt	
Drosophila melanogaster	Y	Unknown protein – conserved	U	NOSIG	AGN10E60-163clu.txt	
Drosophila melanogaster	Y	Unknown protein – conserved across species	U	NoORF	AGN10E60-184clu.txt	

Table 3. *Continued*

Clus[1]	R[2]	GenBank match[3]	E-value[4]	Gi numb[5]	PFAM match[6]	E-value[7]
39	2	No matches found			pfam01028 Topoisomerase_I	2.00E-07
168	1	No matches found			pfam01028 Topoisomerase_I	7.00E-08
232	1	gi|7295456| CG10624 gene product	3.00E-83	gi|7295456	pfam00654 voltage_CLC	4.00E-06

[1]Clus, cluster number.
[2]R, number of sequences of a given custer (R, representation).
[3]GenBank match, best match to the GenBank database.
[4]E-value, indicates significance of match to NR sequence of previous column.
[5]Gi numb, GenBank accession number.
[6]PFAM match, best match to the PFAM database.
[7]E-value, indicates significance of match to CDD sequence of previous column.

cysteine analog in which a selenium atom is found in place of sulphur. Although this family of enzymes has been identified in Bacteria, Archae and Eukarya, being common in mammals (Behne and Kyriakopoulos, 2001), this is the first report of a clone coding for selenoprotein being identified in insects. All the selenoproteins identified thus far are enzymes, with the selenocysteine residue responsible for their catalytic function. Both intracellular and plasma selenoproteins have been identified, indicating that these enzymes are part of the cellular and plasma antioxidant defense system (Behne and Kyriakopoulos, 2001). In fact, pro-oxidants have been involved with processes related to inflammatory reactions, such as endothelial cell injury (Varani and Ward, 1994) and platelet aggregation (Pignatelli et al., 1998); accordingly, we suggest that this putative secreted form of selenoprotein may be involved in attenuation of these reactions. The CLUSTAL alignment of selenoprotein from *A. gambiae* and *D. melanogaster* is shown in Fig. 5.

We have also found a partial-length clone with sequence homology to salivary apyrase and 5′ nucleotidase from *A. gambiae* (Champagne et al., 1995; Arcà et al., 1999). These enzymes play a determinant role in controlling nucleotide concentrations in the blood and preventing platelet aggregation by destroying ADP, a pro-aggregatory molecule necessary for completion of platelet aggregation triggered by most physiological agonists (Francischetti et al., 2000; Gachet, 2000). We have also sequenced a clone with homology to anophelin from *A. albimanus*, a tight-binding inhibitor of thrombin (Francischetti et al., 1999). Calreticulin, selenoprotein, apyrases, 5′ nucleotidase, anophelin and peroxidase (Ribeiro and Valenzuela, 1999) are some of the candidate molecules that provide the redundant anti-hemostatic 'barrier' to prevent host defenses triggered by blood feeding. Finally, we have identified for the first time in the salivary gland of *A. gambiae* a clone coding for a protein similar to *A. aegypti* amylase (sugar digestion) (Grossman and James, 1993) and confirmed the presence of transcripts for lyzozyme in this same tissue (gi:894206) and most likely involved in bacterial cell-wall digestion (Rossignol and Lueders, 1986; Gao and Fallon, 2000).

A clone with a Pfam match for the mucin-like domain has been encountered. This protein has 41 amino acid residues in the signal peptide, which is atypical for the proteins coded by our library; whether this protein is a true full-length clone or a truncated form remains to be determined. Nevertheless, studies on host-pathogen interactions have led to the discovery of various cell surface-associated and secretory mucins. Mucins and mucin-like molecules have recently been described in several protozoan parasites at different life-cycle stages. It is now becoming evident that mucins in parasites are involved in cell-cell interaction and cell surface protection, thus helping the parasite to establish infection (Hicks et al., 1999). Whether *A. gambiae* salivary mucin-like protein could somehow modulate parasite infectivity or help to lubricate insect mouthparts remains to be determined.

Several clusters abundantly expressed in the salivary gland from insects other than *A. gambiae* have been identified in our library. Among these, three cDNA clusters related to proteins of the antigen 5 family (Schreiber et al., 1997) were found. These proteins have been designated here 'antigen 5-related protein 1' (A5R1) (homologous to antigen 5 from *Glossinia morsitans*) (Li et al., 2001), 'antigen 5-related protein 2' (A5R2) (homologous to antigen 5 from *D. melanogaster*) (Megraw et al., 1998) and 'antigen 5-related protein 3' (A5R3) (homologous to antigen 5 from *Lutzomyia longipalpis*) (Charlab et al., 1999). Antigen 5 belongs to the larger CAP family of proteins that has such members as mammal cysteine-rich secretory proteins (crisp), nematode Ag5-Ag3, vespid antigen 5 and plant pathogenesis-related proteins. These secreted proteins share a core sequence of about 200 amino acids whose precise function remains largely unknown. The CLUSTAL alignment of antigen 5-related protein from *A. gambiae*, *G. morsitans*, *D. melanogaster* and *L. longipalpis* is shown in Fig. 6.

We have also found for the first time a cluster with sequence homology to 30 kDa allergen from *A. aegypti* (Brummer-Korvenkotio et al., 1996). Although the precise function of 30 kDa allergen is currently unknown, it is clear that allergic reactions to mosquito bite are of increasing clinical concern. In fact, cutaneous reactions usually

Table 3. *Continued*

Org[8]	N[9]	Function[10]	F[11]	Clone[12]	FASTA[13]	CLUSTAL[14]
	Y	Unknown protein – Topoisomerase domain	U	NoORF	AGN10E60-39clu.txt	AGN10E60-39aln.txt
	Y	Unknown protein Topoisomerase domain		U	NoORF	AGN10E60-168clu.txt
Drosophila melanogaster	Y	Voltage-gated chloride channel?	U	NoORF	AGN10E60-232clu.txt	

[8]Org, organism where a match was found.
[9]N, Novel (Y, yes; N, no).
[10]Function, putative function or biological property.
[11]F, Function; U, unknown
[12]Clone: NOSIG, no signal peptide detected by SignalP server; No ORF, absence of a open-reading frame; ANCH, anchor protein.
[13]FASTA, FASTA-formatted sequences.
[14]CLUSTAL, clustal alignment for two or more sequences.

involving both IgE- and lymphocyte-mediated hypersensitivity are common with insect bites, and systemic reactions including angioedema, generalized urticaria, asthma and anaphylactic shock have been reported (Almeida and Billingsley, 1999; Peng et al., 2001). Accordingly, identification of these potential allergens could lead to their use as markers of bite exposure or, eventually, as antigens for use in immunotherapy (Bousquet et al., 1998). The CLUSTAL alignment of 30 kDa allergen from *A. gambiae* and *A. aegypti* is shown in Fig. 7.

We have also confirmed the presence of previously described D7-related 1–4 transcripts (Arcà et al., 1999) in our library. In addition, a novel full-length D7-related protein containing a typical signal peptide was found, herein designated *A. gambiae* D7-related 5 protein. Furthermore, a novel D7 sequence that codes for a translated mature protein

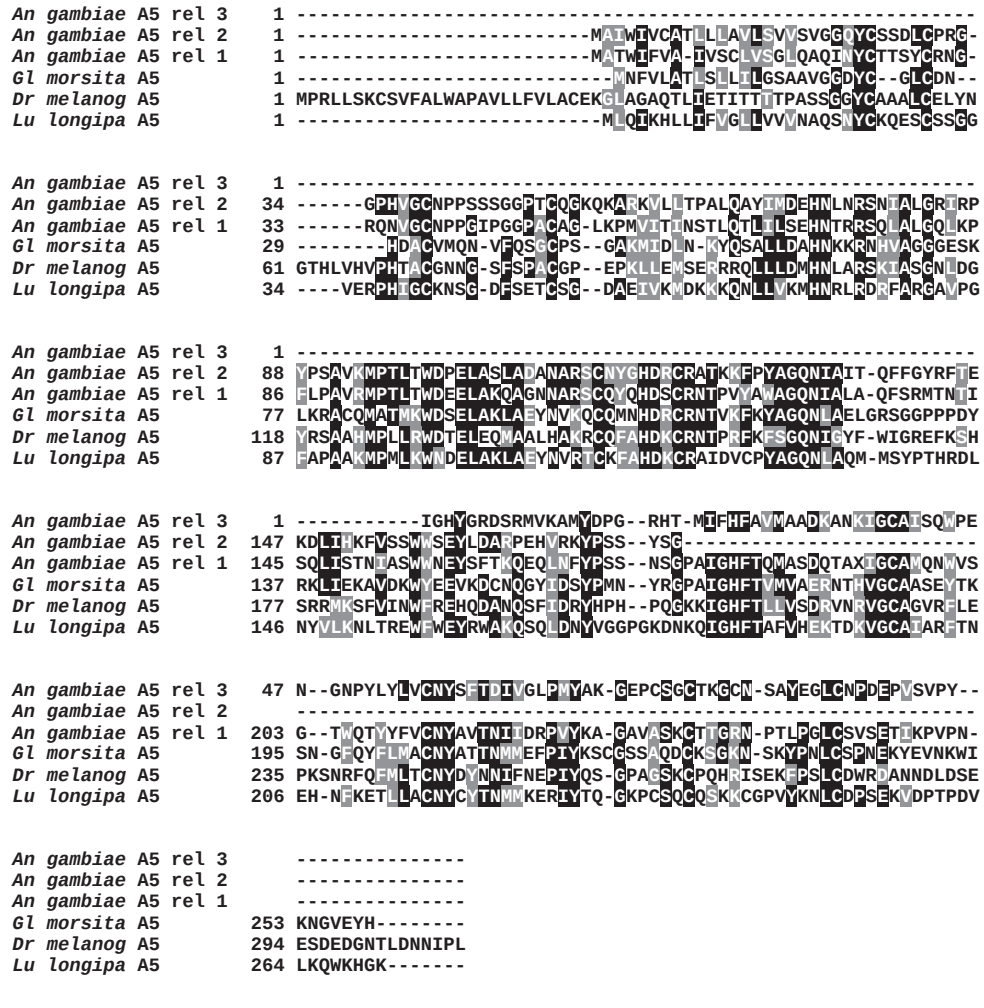

Fig. 6. CLUSTAL alignment of novel *A. gambiae* antigen 5-related proteins (*An gambiae* A5 rel 1, gi 18389882; *An gambiae* A5 rel 2, gi 18389884; and *An gambiae* A5 rel 3, gi 18389886) and *G. morsitans* antigen 5 (*Gl morsita* A5, gi 8927462), *D. melanogaster* antigen 5 (*Dr melanog* A5, gi 7292977), and *L. longipalpis* antigen 5 (*Lu longipa* A5, gi 4887102). Similar amino acid residues are marked with a gray background, identical amino acids with a black background.

of approx. 33 kDa with high similarity to *A. aegypti* D7 protein (James et al., 1991) has also been encountered. A report on the D7 family of salivary proteins in several blood sucking diptera has been recently published (Valenzuela et al., 2002). Accordingly, the putative function of the D7 family is unknown, but the high sequence similarity to odorant-binding proteins suggests that these proteins are carriers for small ligands presumably involved in vector/host interactions (Steinbrecht, 1998). The CLUSTAL alignment for the long form of D7 from *A. gambiae* and *A. aegypti* D7 is depicted in Fig. 8.

A number of other *A. gambiae* sequences have also been reported and they code for the so-called salivary gland (SG) proteins (SG1-8) (Arcà et al., 1999); our library has clones identical to signal peptide-containing SG 1-like 2, SG 2, SG 3, SG 5, SG 6 and SG 7, in addition to SG 1-like proteins herein called SG 1-like 3 and SG 1-like 4, and an SG 7-like molecule herein called SG 7-like 1. We could not identify transcripts for SG 1, SG 4 and SG 8. These proteins have unknown functions. Of interest, nine unique sequences coding for the so-called *A. gambiae* hypothetical proteins (HP) have been reported and designated cE5, c8, c4, c10, c6, A36B, Df2, CB1 and bB2 (Arcà et al., 1999). In our library, we have identified sequences similar or identical to c10, bB2 and cE5. Although cE5 has been designated as a hypothetical protein before, this molecule has been more recently characterized as a potent inhibitor of thrombin (Francischetti et al., 1999). In this regard, we have identified by Edman degradation the N-terminal sequence compatible with CB1; in addition, a bB2-like protein containing the sequence X_6SDSEEA (X_6SDSDEA in bB2) was found (Fig. 1).

Finally, the *A. gambiae* cDNA library has a number of hypothetical proteins characterized by an open reading frame and a putative signal peptide with no database hits.

cDNAs coding for protein that could not be characterized as housekeeping or secretory

Table 3 shows that for a significant number of clones, no significant match to the NR database was found, nor was indication of a signal peptide obtained. Accordingly, these sequences could represent partial housekeeping or secretory cDNA or, alternatively, truncated cDNA.

A catalog for the cDNA from the salivary gland of A. gambiae

To gather the maximum amount of information about the putative secreted proteins from the *A. gambiae* salivary gland, the sequences presented in Table 2 that were classified as 'novel', with or without database hits, were resequenced to obtain, when applicable, their full-length cDNA. The full-coding sequences with database hits were then blasted to the NR protein database and SignalP server to, respectively, confirm sequence novelty and the presence of a signal peptide (Nielsen et al., 1997). In the event a signal peptide was predicted to exist, the molecular mass and the pI of the mature protein were also calculated and, when possible, the function annotated. The same approach was performed for other *A. gambiae* salivary gland cDNAs whose sequences have been reported or deposited in GenBank. The clones without database hits with an open reading frame and a putative signal peptide were subjected to the same bio-informatic analysis and were designated hypothetical proteins (HP), as suggested before (Arcà et al., 1998). In an attempt to provide a uniform and

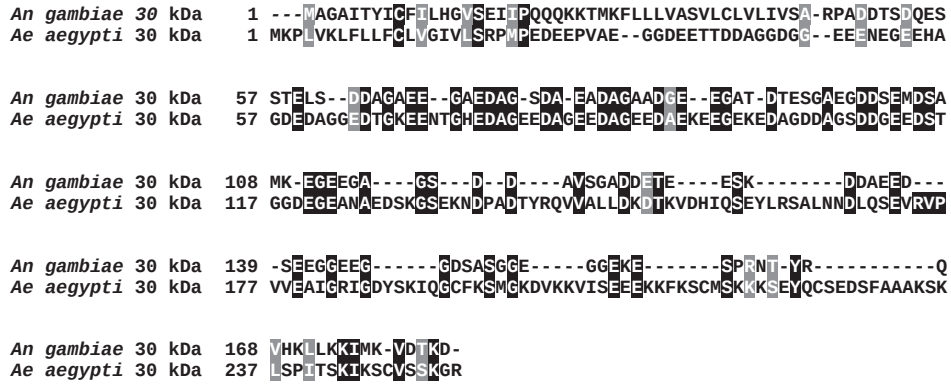

Fig. 7. CLUSTAL alignment of novel *A. gambiae* 30 kDa protein (*An gambiae* 30 kDa; gi 18389878) and *A. aegypti* (*Ae aegypti* 30 kDa, gi 2114497). Similar amino acid residues are marked with a gray background, identical amino acids with a black background.

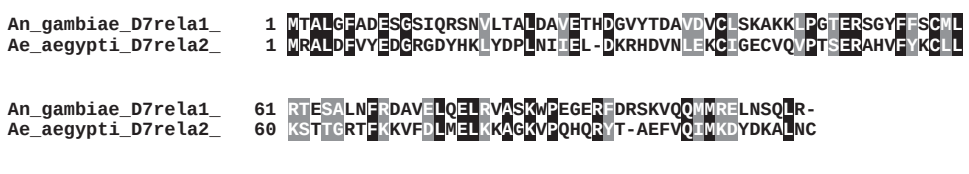

Fig. 8. CLUSTAL alignment of novel *A. gambiae* long D7 protein (*An gambiae* D7 rela1; gi 18389890) and *A. aegypti* long D7 *(Ae aegypti* D7 rela2, gi 159559). Similar amino acid residues are marked with a gray background, identical amino acids with a black background.

comprehensive classification of these hypothetical proteins, we suggest designating each such HP by a given number, beginning with 1. The nine previously described HP have been designated herein HP1–HP9 and the eight novel proteins

Table 4. *A catalog of* Anopheles gambiae *secretory cDNAs**

Sequence name	Clus[1]	Clone[2]	GenBank[3]	N[4]	Mol. Wt.[5]	SP6	Mol. Wt.[7]	pI[8]	Function[9]
30-kDa salivary gland	6	Full-length	gi 18389878	Y	18845.01	45	13957.91	3.59	Antigenic
Amylase	58	Partial	gi 18378722	Y	NA[10]	NA	NA	NA	Digestion
Antigen 5 related 1	14	Full-length	gi 18389882	Y	19682.59	22	17484.81	9.17	Antigenic
Antigen 5 related 2	69	Full-length	gi 18389884	Y	28179.24	19	26187.78	8.74	Antigenic
Antigen 5 related 3	210	Partial	gi 18389886	Y	NA	NA	NA	NA	Antigenic
Apyrase	20	Full-length	gi 4582526	N	61713.67	22	59368.73	8.53	Anti-platelet[12]
Calreticulin	183	Full-length	gi 18389888	Y	46284.78	16	44594.69	4.22	Anti-angiogenic
D7 related, long form	15	Full-length	gi 18389890	Y	35765.59	21	33368.57	6.18	Unknown
D7 related 1, short form	8	Full-length	gi 4538887	N	18665.03	21	16394.2	9.1	Unknown
D7 related 2, short form	1	Full-length	gi 4538889	N	18468.21	21	16166.28	4.75	Unknown
D7 related 3, short form	2	Full-length	gi 4538891	N	18691.51	21	16384.57	4.5	Unknown
D7 related 4, short form	11	Full-length	gi 13537670	N	19330.62	21	16937.61	7.28	Unknown
D7 related 5, short form	26	Full-length	gi 18378062	Y	18872.68	22	18872.68	6.37	Unknown
Hypot. protein 1 (cE5)[11]	17	Full-length	gi 4127344	N	10915.72	21	8681.7	3.75	Anti-coagulant
Hypot. protein 2 (c8)	NA	Full-length	gi 17026156	N	2393.98	16	719.85	3.75	Unknown
Hypot. protein 3 (c4)	NA	Full-length	gi 17026154	N	3745.51	22	1405.6	10.33	Unknown
Hypot. protein 4 (c10)	22	Full-length	gi 13537672	N	20949.88	22	16029.82	8.53	Unknown
Hypot. protein 5 (c6)	NA	Full-length	gi 13509403	N	18859.65	17	17109.4	3.87	Unknown
Hypot. protein 6 (A36B)	NA	Full-length	gi 13509401	N	11171.85	22	8929.13	5.57	Unknown
Hypot. protein 7 (Df2)	NA	Full-length	gi 4127334	N	14068.21	21	11912.57	6.43	Unknown
Hypot. protein 8 (CB1)	NA	Full-length	gi 4127308	N	13490.3	20	11481.82	5.92	Unknown
Hypot. protein 9 (dB2)	7	Full-length	gi 4127306	N	8788.89	18	6837.29	3.97	Unknown
Hypot. protein 10	68	Full-length	gi 18389900	Y	9992.52	23	7517.48	5.5	Unknown
Hypot. protein 11	90	Full-length	gi 18389902	Y	6293.53	25	3336.78	10.48	Unknown
Hypot. protein 12	188	Full-length	gi 18389904	Y	10184.69	21	7974.89	4.52	Unknown
Hypot. protein 13	93	Full-length	gi 18389906	Y	6180.25	22	3802.19	5.49	Unknown
Hypot. protein 14	162	Full-length	gi 18389908	Y	5265.23	17	3308.86	9.78	Unknown
Hypot. protein 15	139	Full-length	gi 18389910	Y	8014.64	30	4859.57	10.56	Unknown
Hypot. protein 16	134	Full-length	gi 18389921	Y	5029.18	27	2210.63	12.1	Unknown
Hypot. protein 17	4	Full-length	gi 18389914	Y	9649.46	43	4912.64	10.91	Unknown
Lysozyme	10	Full-length	gi 894206	N	15396.69	20	13315.05	8.61	Anti-microbial
Mucin-like protein	34	Partial	gi 18389892	Y	NA	NA	NA	NA	Unknown
5′-Nucleotidase	46	Full-length	gi 4582528	N	63458.42	22	61119.55	6.66	Anti-platelet
Salivary gland 1	NA	Full-length	gi 4210615	N	46277.03	16	44229.46	8.04	Unknown
Salivary gland 1-like 2	56	Full-length	gi 13537664	N	43624.38	20	41380.53	7.2	Unknown
Salivary gland 1-like 3	37	Full-length	gi 18389894	Y	30970.37	48	25525.03	5.98	Unknown
Salivary gland 1-like 4	212	Partial	gi 18389896	Y	NA	NA	NA	NA	Unknown
Salivary gland 2	NA	Full-length	gi 4210616	N	11740.16	20	9708.56	3.19	Unknown
Salivary gland 2-like 1	5	Full-length	gi 13537660	N	17137.42	18	15169.9	6.01	Unknown
Salivary gland 3	NA	Full-length	gi 4210618	N	20037.83	18	18248.55	4.62	Unknown
Salivary gland 5	13	Full-length	gi 13537662	N	38201.55	24	35802.65	6.27	Unknown
Salivary gland 6	9	Full-length	gi 13537666	N	13094.35	28	10089.52	5.61	Unknown
Salivary gland 7	12	Full-length	gi 13537668	N	16327.77	25	13749.58	8.15	Unknown
Salivary gland 7-like 1	19	Full-length	gi 18389898	Y	10208.16	25	7419.68	10.26	Unknown
Salivary gland 8	NA	Full-length	gi 13537673	N	14297.76	24	11684.5	10.16	Unknown
Selenoprotein	128	Full-length	gi 18389880	Y	18424.37	19	16373.76	4.89	Anti-oxidant

*Including putative secretory proteins.
[1]Clus, cluster number.
[2]Clone, clone type (partial or full-length).
[3]GenBank, NR database accession number.
[4]N, novel (Y=yes; N=No).
[5]Mol. Wt., molecular mass before signal peptide removal.
[6]SP, signal peptide.
[7]Mol. Wt., molecular mass after signal peptide removal.
[8]pI, Isoelectric point.
[9]Function, putative function or biological properties.
[10]NA, not available.
[11]Hypot. protein, Hypothetical protein.
[12]For function annotation, references are given in the text.

described in this paper have been named HP10–HP17 (see Table 4). Formal characterization of such proteins and their biological function remains to be determined.

Taking into account the 21 novel *A. gambiae* sequences described herein in addition to 25 previously described, there are 46 different salivary gland cDNAs coding for putative secreted proteins, most of them (42 sequences) being full-length clones with a clear signal peptide. The four remaining partial clones have been classified as secretory, based on their high sequence similarity to other unambiguously characterized extracellular proteins. Accordingly, SG 4 was not included in Table 4 since no signal peptide could be detected for this protein. Interestingly, we have found in our library the cDNAs corresponding to most proteins whose amino terminus had previously been identified by Edman degradation (Fig. 1, Table 4). In contrast, and as expected, the amino terminus of many putative proteins coded by secretory cDNA shown in Table 4 could not be identified, either because the protein is expressed in low-copy number or because of technical limitations inherent to Edman degradation. In some cases, the apparent molecular mass of some proteins detected by SDS-PAGE (Fig. 1) is different from that predicted by the cDNA (Table 4). This is most likely due to protein glycosylation or formation of dimers that have not been appropriately separated by SDS.

To our knowledge, Table 4 is the first attempt to create a comprehensive catalog of the cDNAs from the *A. gambiae* salivary gland coding for putative secretory proteins. It is clear from this set of cDNAs that many proteins could not have their putative function annotated. Eventually, however, such a catalog will contain a nonredundant set of full-coding cDNA sequences covering every *A. gambiae* salivary gland cDNA and possibly each salivary protein function. Thus, this transcript and protein catalog could form part of a large-scale and comprehensive functional analysis of mosquito genes and, together with information derived from *Plasmodium* spp. genome, could be an essential tool for understanding the molecular basis of malaria.

We are grateful to Drs Robert W. Gwadz, Thomas J. Kindt and Louis H. Miller for encouragement and support. We also thank Brenda Rae Marshall for editorial support.

References

Altland, K. (1990). IPGMAKER: a program for IBM-compatible personal computers to create and test recipes for immobilized pH gradients. *Electrophoresis* **11**, 140–147.

Adam, D. (2001). Gene sequencers hope to put the bite on mosquitoes. *Nature* **410**, 137.

Almeida, A. P. G. and Billingsley, P. F. (1999). Induced immunity against the mosquito *Anopheles stephensi*: reactivity characteristics of immune sera. *Med. Vet. Ent.* **13**, 53–64.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389–3402

Arcà, B., Lombardo, F., Capurro, M. L., Torre, A., Dimopouolos, 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. USA* **96**, 1516–1521.

Balter, M. (2001). Sequencing set for dreaded mosquito. *Science* **291**, 1873.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L. and Sonnhammer, E. L. (2000). The PFAM protein families database. *Nucl. Acids Res.* **28**, 263–266.

Behne, D. and Kyriakopoulos, A. (2001). Mammalian selenium-containing proteins. *Ann. Rev. Nutr.* **21**, 453–473.

Bjellqvist, B., Basse, B., Olsen, E. and Celis, J. E. (1994). Reference points for comparison of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **15**, 529–539.

Bousquet, J., Lockey, R., Malling, H. J., Alvarez-Cuesta, E., Canonica, G. W., Chapman, M. D., Creticos, P. J., Dayer, J. M., Durham, S. R., Demoly, P. et al. (1998). Allergen immunotherapy: therapeutic vaccines for allergic diseases. *Ann. Allergy Asthma Immunol.* **81**, 401–405.

Bowman, S., Lawson, D., Basham, D., Brown, D., Chillingworth, T., Churcher, C. M., Craig, A., Davies, R. M., Devlin, K., Feltwell, T. et al. (1999). The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. *Nature* **400**, 532–538.

Brummer-Korvenkotio, H., Palosuo, T., François, G. and Reunala, T. (1996). Characterization of *Aedes communis, Aedes aegypti* and *Anopheles stephensi* mosquito saliva antigens by immunoblotting. *Int. Arch. Allergy Immunol.* **112**, 169–174.

Carlton, J. M., Muller, R., Yowell, C. A., Fluegge, M. R., Sturrock, K. A., Pritt, J. R., Vargas-Serrato, E., Galinski, M. R., Barnwell, J. W., Mulder, N., Kanapin, A., Cawley, S. E., Hide, W. A. and Dame, J. B. (2001). Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. *Mol. Biochem. Parasitol.* **118**, 201–210.

Cerami, C., Frevert, U., Sinnis, P., Takacs, B., Clavijo, P., Santos, M. J. and Nussenzweig, V. (1992). The basolateral domain of the hepatocyte plasma membrane bears receptors for the circumsporozoite protein of *Plasmodium falciparum* sporozoites. *Cell* **70**, 1021–1033.

Champagne, D. E., Smartt, C. T., Ribeiro, J. M. and James, A. A. (1995). The salivary gland-specific apyrase of the mosquito *Aedes aegypti* is a member of the 5′-nucleotide family. *Proc. Natl. Acad. Sci. USA* **92**, 694–699.

Charlab, R., Valenzuela, J. G., Rowton, E. D. and Ribeiro, J. M. C. (1999). Toward an understanding of the biochemical and pharmacological complexity of the saliva of a hematophagous sand fly *Lutzomyia longipalpis*. *Proc. Natl. Acad. Sci. USA* **96**, 15155–15160.

Collins, F. H. and Paskewitz, S. M. (1995). Malaria: current and future prospects for control. *Annu. Rev. Entomol.* **40**, 195–219.

Delanghe, J. R. and Langlois, M. R. (2001). Hemopexin: a review of biological aspects and the role in laboratory medicine. *Clin. Chim. Acta* **312**, 13–23.

Dragoni, I., Mariotti, M., Consalez, G. G., Soria, M. R. and Maier, J. A. (1998). EDF-1, a novel gene product down-regulated in human endothelial cell differentiation. *J. Biol. Chem.* **273**, 31119–31124.

Fauci, A. S. (2001). Infectious diseases: considerations for the 21st century. *Clin. Infect. Dis.* **32**, 675–685.

Francischetti, I. M. B., Ribeiro, J. M. C., Champagne, D. and Andersen, J. (2000). Purification, cloning, expression, and mechanism of action of a novel platelet aggregation inhibitor from the salivary gland of the blood-sucking bug, *Rhodnius prolixus*. *J. Biol. Chem.* **275**, 12639–12650.

Francischetti, I. M. B., Valenzuela J. G. and Ribeiro, J. M. (1999). Anophelin: kinetics and mechanism of thrombin inhibition. *Biochemistry* **38**, 16678–16685.

Gachet, C. (2000). Platelet activation by ADP: the role of ADP antagonists. *Ann. Med.* **32**, 15–20.

Gao, Y. and Fallon, A. M. (2000). Immune activation upregulates lysozyme gene expression in *Aedes aegypti* mosquito cell culture. *Insect Mol. Biol.* **9**, 553–558.

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C. et al. (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **282**, 1126–1132.

Griffioen, A. W. and Molema, G. (2000). Angiogenesis: potentials for pharmacologic intervention in the treatment of cancer, cardiovascular diseases, and chronic inflammation. *Pharmacol. Rev.* **52**, 237–268.

Grossman, G. L. and James, A. A. (1993). The salivary glands of the vector mosquito, *Aedes aegypti*, express a novel member of the amylase gene family. *Insect Mol. Biol* **1**, 223–232.

Henikoff, S. and Henikoff, J. G. (1994). Protein family classification based on searching a database of blocks. *Genomics* **19**, 97–107.

Hicks, S. J., Theodopoulos, G., Carrington, S. D. and Corfield, A. P. (1999). The role of mucins in host–parasite interactions. I. Protozoan parasites. *Parasitol. Today* **16**, 476–481.

Higgins, D. G., Thompson, J. D. and Gibson, T. J. (1996). Using CLUSTAL or multiple sequence alignments. *Methods Enzymol.* **266**, 383–402.

Holmgren, A. and Bjornstedt, M. (1995). Thioredoxin and thioredoxin reductase. *Methods Enzymol.* **52**, 199–208.

Huribut, H. S. (1966). Mosquito salivation and virus transmission. *Am. J. Trop. Med. Hyg.* **15**, 989–993.

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.

Janssen, C. S., Barrett, M. P., Lawson, D., Quail, M. A., Harris, D., Bowman, S., Phillips, R. S. and Turner, C. M. R. (2001). Gene discovery in *Plasmodium chabaudi* by genome survey sequencing. *Mol. Biochem. Parasitol.* **113**, 251–269.

Johnson, S., Michalak, M., Opas, M. and Eggleton, P. (2001). The ins and outs of calreticulin: from the ER lumen to the extracellular space. *Trends Cell Biol.* **11**, 122–129.

Kappe, S. H. I., Gardner, M. J., Brown, S. M., Ross, J., Matuschewski, K., Ribeiro, J. M., Adams, J. H., Quackenbush, J., Cho, J., Carucci, D. J., Hoffman, S. L. and Nussenzweig, V. (2001). Exploring the transcriptome of the malaria sporozoite stage. *Proc. Natl. Acad. Sci. USA* **98**, 9895–9900.

Krettli, A. U. and Miller, L. H. (2001). Malaria: a sporozoite runs through it. *Curr. Biol.* **11**, 409–412.

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

Lin, M. Z. and Greenberg, M. E. (2000). Orchestral maneuvers in the axon: TRIO and the control of axon guidance. *Cell* **101**, 239–242.

Lund, P. A. (2001). Microbial molecular chaperones. *Adv. Microb. Physiol.* **44**, 93–140.

Miao, W. M., Vasile, E., Lane, W. S. and Lawler, J. (2001). CD36 associates with CD9 and integrins on human blood platelets. *Blood* **97**, 1689–1696.

Megraw, T., Kaufman, T. C. and Kovaick, G. E. (1998). Sequence and expression of *Drosophila* antigen 5-related 2, a new member of the CAP gene family. *Gene* **19**, 297–304.

Nakamura, H., Nakamura, K. and Yodoi, J. (1997). Redox regulation of cellular activation. *Annu. Rev. Immunol.* **15**, 351–369.

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.

Peng, Z., Xu, W., James, A. A., Lam, H., Sun, D., Cheng, L. and Simons, F. E. (2001). Expression, purification, characterization and clinical relevance of rAed a 1-a 68-kDa recombinant mosquito *Aedes aegypti* salivary allergen. *Int. Immunol.* **13**, 1445–1452.

Pike, S. E., Yao, L., Jones, K. D., Cherney, B., Appella, E., Sakaguchi, K., Nakhasi, H., Teruya-Feldstein, J., Wirth, P., Gupta, G. and Tosato, G. (1998). Vasostatin, a calreticulin fragment, inhibits angiogenesis and suppresses tumor growth. *Blood* **188**, 2349–2356.

Pignatelli, P., Pulcinelli, F. M., Lenti, L., Gazzaniga, P. P. and Violi, F. (1998). Hydrogen peroxide is involved in collagen-induced platelet activation. *Blood* **91**, 484–490.

Ribeiro, J. M. C. (1987). Role of blood feeding arthropods in blood feeding. *Ann. Rev. Entomol.* **32**, 463–478.

Ribeiro, J. M. C. and Valenzuela, J. G. (1999). Purification and cloning of the salivary peroxidase/catechol oxidase of the mosquito *Anopheles albimanus. J. Exp. Biol*. **202**, 809–816.

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

Sibbald, P. R., Sommerfeldt, H. and Argos, P. (1991). Automated protein sequence pattern handling and PROSITE searching. *Comp. Appl. Biosci.* **7**, 535–536.

Schreiber, M. C., Karlo, J. C. and Kovalick, G. E. (1997). A novel cDNA from *Drosophila* encoding a protein with similarity to mammalian cysteine-rich secretory proteins, wasp venom antigen 5, and plant group 1 pathogenesis-related proteins. *Gene* **191**, 135–141.

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

Sidjanski, S. and Vanderberg, J. R. (1997). Delayed migration of plasmodium sporozoites from the mosquito bite site to the blood. *Am. J. Trop. Med. Hyg.* **57**, 426–429.

Steinbrecht, R. A. (1998). Odorant-binding proteins: expression and function. *Ann. NY Acad. Sci.* **30**, 323–332.

Takemaru, K.-I., Li, F.-Q., Ueda, H. and Hirose, S. (1997). Multiprotein bridging factor 1 (MBF1) is an evolutionarily conserved transcriptional coactivator that connects a regulatory factor and TATA element-binding protein. *Proc. Natl. Acad. Sci. USA* **94**, 7251–7256.

Todres, E., Nardi, J. B. and Robertson, H. M. (2000). The tetraspanin superfamily in insects. *Insect Mol. Biol.* **9**, 581–590.

Touray, M. G., Warburg, A., Laughinghouse, A., Krettli, A. U. and Miller, L. H. (1992). Developmentally regulated infectivity of malaria sporozoites for mosquito salivary glands and the vertebrate host. *J. Exp. Med.* **175**, 1607–1612.

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

Varani, J. and Ward, P. A. (1994). Mechanism of endothelial cell injury in acute inflammation. *Shock* **2**, 311–319.

Young, J. C., Moarefi, I. and Hartl, F. U. (2001). Hsp90: a specialized but essential protein-folding tool. *J. Cell Biol.* **154**, 267–273.