REGULAR ARTICLE

A proteomic analysis of salivary glands of female *Anopheles gambiae* mosquito

Dário E. Kalume[1], Mobolaji Okulate[2], Jun Zhong[1], Raghunath Reddy[1], Shubha Suresh[3], Nandan Deshpande[3], Nirbhay Kumar[2] and Akhilesh Pandey[1]*

[1] McKusick-Nathans Institute of Genetic Medicine and Department of Biological Chemistry, Johns Hopkins School of Medicine, Baltimore, MD, USA
[2] Department of Molecular Microbiology and Immunology, Johns Hopkins Malaria Research Institute, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[3] Institute of Bioinformatics, Discoverer Unit 1, Bangalore, India

Received: September 3, 2004
Revised: December 16, 2004
Accepted: December 27, 2004

Understanding the development of the malaria parasite within the mosquito vector at the molecular level should provide novel targets for interrupting parasitic life cycle and subsequent transmission. Availability of the complete genomic sequence of the major African malaria vector, *Anopheles gambiae*, allows discovery of such targets through experimental as well as computational methods. In the female mosquito, the salivary gland tissue plays an important role in the maturation of the infective form of the malaria parasite. Therefore, we carried out a proteomic analysis of salivary glands from female *An. gambiae* mosquitoes. Salivary gland extracts were digested with trypsin using two complementary approaches and analyzed by LC-MS/MS. This led to identification of 69 unique proteins, 57 of which were novel. We carried out a functional annotation of all proteins identified in this study through a detailed bioinformatics analysis. Even though a number of cDNA and Edman degradation-based approaches to catalog transcripts and proteins from salivary glands of mosquitoes have been published previously, this is the first report describing the application of MS for characterization of the salivary gland proteome. Our approach should prove valuable for characterizing proteomes of parasites and vectors with sequenced genomes as well as those whose genomes are yet to be fully sequenced.

Keywords:
Anopheles gambiae / Malaria / MS

1 Introduction

Malaria remains a leading cause of morbidity and mortality in the tropical and sub-tropical regions of the world and is responsible for at least one million deaths annually [1]. Emergence of resistance in the *Plasmodium* parasites and mosquito vectors combined with poor knowledge of mosquito biology and inappropriate vector control strategies has hindered numerous attempts to combat this disease [2, 3]. About 60 *Anopheles* mosquito species have been shown to transmit the human malaria worldwide. Among them, *An. gambiae* is the most competent and the primary vector in sub-Saharan Africa [4].

A critical first step during the transmission of malaria is the ingestion of *Plasmodium* gametocytes into the midgut of the female mosquito during a blood meal. After a

Correspondence: Dr. Akhilesh Pandey, McKusick-Nathans Institute of Genetic Medicine and Department of Biological Chemistry, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA
E-mail: pandey@jhmi.edu
Fax: +1-410-502-7544

Abbreviations: OBP, odorant binding protein; **SG-like**, salivary gland-like

* Additional corresponding author: Dr. Nirbhay Kumar, **E-mail:** nkumar@jhsph.edu

period of sexual development in the midgut, sporozoites, the infective form of the parasite, are produced, which then migrate to the salivary glands and are transmitted to its vertebrate host while feeding. It has been hypothesized that cell surface molecules in the salivary glands of female mosquitoes play a critical role in the transmission of malaria parasite. However, to date, limited studies have characterized molecular interactions between the salivary glands of the mosquito and the sporozoites of the *Plasmodium* parasite. It has been previously shown that sporozoite-salivary gland interaction is species specific and receptor mediated; the glands being involved in both recognition and invasion [5]. It is also known that sporozoites only invade the distal parts of median and lateral lobes of female salivary glands and that recombinant circumsporozoite (CS) protein binds specifically to *Anopheles stephensi* salivary glands, particularly to the median and distal lateral lobes of the gland [6]. Mosquito saliva contains a large number of biomolecules responsible for antihemostatic activity, which assist hematophagous arthropods during the feeding process [7].

The recent completion of *An. gambiae* genome sequence [8] provided an architectural scaffold for mapping, identifying, selecting, and exploiting desirable insect vector genes. The annotation of the *An. gambiae* genome sequence has been an ongoing process since it was completed in 2002 [8]. The assembled genome is publicly available through NCBI (National Center for Biotechnology Information) and EMBL-EBI (European Bioinformatics Institute)/Ensembl (http://www.ensembl.org). Here, we provide a proteomic analysis of adult female *An. gambiae* salivary glands. We anticipate that further elucidation of the novel protein targets identified in this study will shed more light on the biology of malaria transmission and perhaps suggest novel targets for control of malaria transmission.

2 Materials and methods

2.1 Reagents

Vydac C_{18} resin was provided by Nest Group (Southboro, MA, USA) and YMC gel ODS-A from YMC (Kyoto, Japan). Sequencing grade modified trypsin was purchased from Promega (Madison, WI, USA). For protein quantitation Bio-Rad DC Protein assay kit was supplied from Bio-Rad Laboratories (Hercules, CA, USA). DTT and iodoacetamide (IAA) were from Fluka Chemie (Buchs, Switzerland) and Sigma (St.Louis, MO, USA), respectively. All solutions were prepared with ultra high purity Milli-Q water.

2.2 Mosquitoes and salivary gland isolation

An. gambiae mosquitoes (G-3 strain) initially obtained from the Laboratory of Parasitic Diseases (National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda) were maintained in humidity and temperature controlled insectaries at Johns Hopkins University. *An. gambiae* mosquitoes were reared at 27 ± 1°C and 80 ± 5% relative humidity with 12-h cycles of alternating darkness and light. Adult mosquitoes were maintained on 10% Karo Dark Corn Syrup, except before harvesting the salivary glands. The female adult mosquitoes were allowed to feed on malaria parasite-free human blood for 30 min, and subsequently rested overnight. The salivary glands (one pair *per* individual) were then dissected and immediately placed in PBS and then stored at −70°C.

2.3 Sample preparation and electrophoresis

Proteins were extracted from 100 pairs of salivary glands by homogenization of the tissue in PBS using ultrasonication (Sonifier cell disruptor, Branson, CT, USA) followed by three freeze-thaw cycles. The extracted suspension was centrifuged for 20 min 14 000 rpm at 4°C, and the supernatant collected for in-solution digestion with trypsin. Another 150 pairs of dissected salivary glands were lysed in buffer containing 50 mM Tris-HCl pH 7.4, 150 mM NaCl and 1% NP-40, containing protease inhibitors (Complete protease inhibitor cocktail tablets; Roche Diagnostics, Mannheim, Germany), followed by ultrasonication. The lysate was then fractionated by SDS-PAGE. The gel was stained with colloidal CBB according to the manufacturer's protocol (NuPAGE Novex; Invitrogen, Carslbad, CA, USA). Following staining, the gel was sliced into 15 bands (approximately 1 cm^2 each) and the gel slices subjected to digestion with trypsin.

2.4 Trypsin digestion

2.4.1 In-gel digestion

Approximately 23 μg homogenate derived from 90 pairs of salivary glands was loaded onto an SDS-PAGE gel. The gel was stained with colloidal CBB and the bands digested essentially as described previously [9]. Briefly, the gel slices were rinsed twice in water, and water : methanol (1:1) solution, dehydrated with ACN, then 10 mM DTT for 45 min at 56°C, followed by 55 mM IAA for 30 min at 25°C to reduce and alkylate the proteins, respectively. Gel slices were then incubated with trypsin solution (6.5 ng/μL), in a final volume sufficient to cover the gels, overnight at 37°C. The resulting peptide mixture was analyzed by LC-MS/MS.

2.4.2 In-solution digestion

The extracted protein was quantified according to a modified Lowry method (Bio-Rad DC Protein assay) and 25 μg sample was used for digestion. Following denaturation with

4 M urea, the proteins were reduced with 10 mM DTT for 40 min at 60°C under nitrogen. Subsequently, the reduced proteins were alkylated with 25 mM IAA for 30 min at 25°C. Subsequently, 100 mM ammonium bicarbonate buffer pH 8.1 was added to the protein mixture solution to dilute urea to a final concentration of 0.5 M. The protein solution was then incubated with trypsin in an enzyme:substrate ratio of 1:50 overnight at 37°C.

2.5 MS analysis

The samples from either in-gel or in-solution digestions were loaded online onto a fused silica capillary column in tandem with a pre-column packed with 5-μm Vydac C_{18} resin and with 12-μm YMC gel ODS-A, respectively. The peptides derived from in-gel digestion were separated using a linear gradient elution from 87% mobile phase A (100% H_2O with 0.4% acetic acid and 0.005% heptafluorobutyric acid) to 40% mobile phase B (90% ACN with 10% H_2O, 0.4% acetic acid and 0.005% heptafluorobutyric acid) in 34 min. The peptides derived from in-solution digestion were separated using a longer linear gradient elution from 88% mobile phase A to 45% mobile phase B in 84 min. A potential of 2.5 kV was applied to the emitter in the ion source. The spectra were acquired on a Micromass Q-TOF API-US (Manchester, UK) equipped with an ion source sample introduction system designed at Proxeon Biosystems (Odense, Denmark). The acquisition and the deconvolution of data were performed on a MassLynx Windows NT PC data system (version 4). All spectra were obtained in the positive-ion mode.

2.6 Data analysis

Data analysis pipeline for peptide and protein identification is described below and is in accordance with the Molecular and Cellular Proteomics editorial board [10]. Mass-Lynx was employed to generate peak lists (pkl files) from the raw data using the following parameters: (a) smooth windows (channels): 4.00, number of smooths: 2, smooth mode: Savitzky Golay; (b) percentage of peak height to calculated the centroid spectra, 80%; and (c) no baseline subtract was allowed. The processed MS/MS spectra were searched against the non-redundant protein database and the *An. gambiae* protein database downloaded from NCBI website (ftp://ftp.ncbi.nih.gov/genbank/genomes/Anopheles_gambiae/). MS data searches were performed using MASCOT version 1.9 [11] installed on a Linux cluster. The following settings were used: (a) trypsin as the specific enzyme (allow up to 2 missed cleavages); (b) peptide window tolerance (error window on experimental peptide mass values) ±0.4 Da; and (c) fragment mass tolerance of ±0.3 Da. Moreover, during the searches, oxidation of methionine and carbamidomethylcysteine modification were the two amino acid modifications allowed. The assignments by MASCOT were verified by manual interpretation of the spectra. In general, only peptides (mass spectra) with a MASCOT score above 30 and containing a sequence tag of at least four consecutive amino acids were considered in this study. Otherwise, mass spectra with lower score, but presenting a clear tandem mass spectrum, were manually interpreted. Sequence coverage and the peptide sequences that match each identified protein are shown in parentheses in Tables. 1–4 and in Supplementary Table 1, respectively. A domain analysis was conducted for all the proteins identified in this study by subjecting the sequences to SMART (http://smart.embl-heidelberg.de/) [12]. Protein sequences were run through a BLASTP search against the non-redundant database to find homologs for annotation. Wherever possible, putative biological processes and localizations were assigned. In this article, novel proteins correspond to those entries whose transcripts, genes or peptide sequences are not described in the literature or found in the publicly available NCBI non-redundant protein database as a named entity.

Table 1. A list of known proteins identified by MS using the gel-free approach

	Name of protein[a), b)]	Accession number	Domains/motifs
1.	D7-related 1 protein (63%)	gi|31222471	Pheromone/OBP
2.	D7-related 2 protein (81%)	gi|4538889	Pheromone/OBP
3.	D7-related 3 protein (42%)	gi|4538891	Pheromone/OBP
4.	D7r4 protein (47%)	gi|13537670	Pheromone/OBP
5.	D7-related 5 protein (15%)	gi|18378603	No conserved domains
6.	**Putative gVAG protein precursor (51%)**	**gi|31217598**	SCP-like extracellular protein
7.	Histone H3 (27%)	gi|1731925	Histones H3 and H4
8.	gSG6 protein (55%)	gi|13537666	No conserved domains
9.	**TRIO protein (20%)**	**gi|18389917**	No conserved domains

a) Proteins found both by in-gel and in-solution approaches are shown in bold
b) The percentage of sequence covered by identified peptides is indicated in parentheses

Table 2. A list of novel proteins identified by MS using the gel-free approach

	Accession number[a), b)]	Domains/motifs	Features
1.	**gi|31222536 (40%)**	Pheromone/OBP	Orthologous to D7 protein in *An. stephensi*
2.	gi|31222545 (51%)	Pheromone/OBP	Similar to D7 protein long form
3.	**gi|31239469 (5%)**	Protein disulfide isomerase	Similar to protein disulfide-isomerase in *D. melanogaster*
4.	**gi|31198983 (2%)**	Protein disulfide isomerase	Similar to protein disulfide-isomerase in *D. melanogaster*
5.	**gi|31205001 (6%)**	Animal haem peroxidase	Orthologous to salivary peroxidase in *Anopheles albimanus*
6.	**gi|31197357 (26%)**	5′-nucleotidase/2′,3′-cyclic phosphodiesterase and related esterases	Similar to putative 5′-nucleotidase (gi|4582528)
7.	gi|31199067 (9%)	C-type lysozyme and alpha-lactalbumin	Similar to lysozyme precursor (EC 3.2.1.17)
8.	**gi|31208237 (3%)**	HSP70 superfamily	Similar to heat shock 70 kDa protein cognate in *Bombyx mori*
9.	gi|31196575 (54%)	No conserved domains	Similar to gSG7 protein
10.	**gi|31203175 (15%)**	No conserved domains	Similar to gSG1b protein
11.	gi|4127307 (19%)	No conserved domains	Hypothetical protein (gi|31217903)
12.	gi|31222305 (19%)	No conserved domains	Hypothetical protein 12 (gi|18389905)
13.	gi|31203045 (12%)	No conserved domains	No significant similarity to other proteins
14.	**gi|31203049 (42%)**	No conserved domains	Long form of the misannotated protein SG1-like 3 protein
15.	gi|31234764 (54%)	No conserved domains	Similar to 30-kDa protein
16.	**gi|18873404 (11%)**	No conserved domains	Hypothetical protein (*An. gambiae*)
17.	gi|18389915 (17%)	No conserved domains	Hypothetical protein 17 (*An. gambiae*)

a) Proteins found both by in-gel and in-solution approaches are shown in bold
b) The percentage of sequence covered by identified peptides is indicated in parentheses

Table 3. A list of known proteins identified by MS using the in-gel digestion approach

	Name or accession number (EnsEMBL) of predicted protein[a), b)]	Accession number	Domains/motifs
1.	Calreticulin (13%)	gi|18389889	Calreticulin
2.	D7 protein long form (51%)	gi|18389891	Pheromone/OBP
3.	**Putative gVAG protein precursor (51%)**	gi|31217598	SCP-like extracellular protein
4.	SG1 protein (22%)	gi|4210615	No conserved domains
5.	**TRIO protein (48%)**	gi|18389917	No conserved domains

a) Proteins found both by in-gel and in-solution approaches are shown in bold
b) The percentage of sequence covered by identified peptides is indicated in parentheses

Table 4. A list of novel proteins identified by MS using the in-gel digestion approach

	Accession number[a), b)]	Domains/motifs	Features
1.	gi|31241043 (11%)	Alpha-amylase	Orthologous to maltase precursor in *Aedes aegypti*
2.	**gi|31205001 (5%)**	Animal haem peroxidase	Orthologous to salivary peroxidase in *A. albimanus*
3.	gi|31201635 (8%)	Contains tubulin domain	Beta tubulin
4.	gi|31228364 (1%)	Contains Kelch motif and fibronectin type 3 domain	No significant similarity to other proteins
5.	gi|21292024 (3%)	Fibrinogen-related domains	No significant similarity to other proteins
6.	gi|31198555 (20%)	14-3-3 family (multifunctional chaperone)	Similar to 14-3-3 protein in *D. melanogaster*
7.	gi|31248144 (16%)	14-3-3 family	Similar to CG31196-PC in *D. melanogaster*

Table 4. Continued

	Accession number[a), b)]	Domains/motifs	Features
8.	gi|31210101 (12%)	ATP synthase alpha/beta family	Similar to ATP synthase, H^+ transporting, mitochondrial F1 complex, alpha subunit, isoform 1, cardiac muscle in *Homo sapiens*
9.	gi|31240019 (18%)	F0F1-type ATP synthase, beta subunit	Similar to ATP synthase-beta CG11154-PA in *D. melanogaster*
10.	gi|31232142 (10%)	Glyceraldehyde-3-phosphate dehydrogenase/ erythrose-4-phosphate dehydrogenase	Similar to glyceraldehyde-3-phosphate dehydrogenase in *Plutella xylostella*
11.	gi|31208673 (8%)	Ribosomal protein L3	Similar to ribosomal protein L3 CG4863-PA in *D. melanogaster* subunit
12.	gi|31201015 (3%)	Translation initiation factor 2, alpha	Similar to eIF2 alpha subunit in *Spodoptera frugiperda*
13.	gi|31209765 (15%)	Gamma-interferon inducible lysosomal thiol reductase	No significant similarity (above 70% identity) to other proteins
14.	**gi|31239469 (17%)**	Protein disulfide isomerase	Similar to protein disulfide-isomerase in *D. melanogaster*
15.	**gi|31198983 (12%)**	Protein disulfide isomerase	Similar to protein disulfide-isomerase in *D. melanogaster*
16.	gi|31241427 (8%)	Protein disulfide isomerase	No significant similarity to other proteins
17.	gi|31208765 (9%)	60S acidic ribosomal protein P0 and ribosomal protein L-10	Orthologous to ribosomal protein P0 in *Aedes albopictus*
18.	gi|31201059 (7%)	60S ribosomal protein L11	Similar to 60S ribosomal protein L11 in *D. melanogaster*
19.	gi|31204623 (3%)	Vacuolar H^+-ATPase V0 sector, subunit d	Similar to vacuolar ATP synthase subunit d 1 in *D. melanogaster*
20.	gi|31207751 (10%)	Vacuolar H^+-ATPase V1 sector, subunit A	Similar to vacuolar ATP synthase catalytic subunit A in *H. sapiens*
21.	**gi|31208237 (22%)**	HSP70 superfamily	Similar to heat shock 70 kDa protein cognate in *B. mori*
22.	gi|31241095 (12%)	HSP70 superfamily	Similar to heat shock cognate 70 in *Chironomus tentans*
23.	gi|31242551 (21%)	Histidine kinase-like ATPases and HSP90 family	Similar to glycoprotein 93 CG5520-PA in *D. melanogaster*
24.	gi|31199877 (3%)	Contains S4 RNA-binding domain	Similar to ribosomal protein S4 in *S. frugiperda*
25.	gi|31240295 (22%)	5′-nucleotidase/2′,3′-cyclic phosphodiesterase and related esterases	Similar to apyrase
26.	**gi|31197357 (25%)**	5′-nucleotidase/2′,3′-cyclic phosphodiesterase and related esterases	Similar to putative 5′-nucleotidase
27.	gi|31203139 (9%)	ERM, Ezrin/radixin/moesin family	Similar to moesin-like CG10701-PA *D. melanogaster*
28.	gi|31203141 (5%)	ERM, Ezrin/radixin/moesin family	Similar to moesin in *D. melanogaster*
29.	gi|31214711 (6%)	TRAP-beta, Translocon-associated protein beta	No significant similarity to other proteins
30.	**gi|31222536 (58%)**	Pheromone/OBP	Orthologous to D7 protein in *An. stephensi*
31.	gi|31198963 (10%)	Translation elongation factor EF-1alpha	Similar to elongation factor 1 alpha *B. mori*
32.	gi|31241317 (4%)	20S proteasome	Similar to proteasome alpha7 subunit CG1519-PA in *D. melanogaster*
33.	gi|31226204 (4%)	Enolase	Similar to enolase CG17654-PB in *D. melanogaster*
34.	gi|31206155 (8%)	Archaeal/vacuolar-type H^+-ATPase subunit	Orthologous to vacuolar ATPase B subunit in *Aedes aegypti*
35.	gi|31204457 (22%)	Contains actin domain	Similar to actin 5C CG4027-PB in *D. melanogaster*
36.	gi|31240645 (8%)	40S ribosomal protein SA (P40)/laminin receptor 1	Similar to ribosome-associated protein P40 in *B. mori*
37.	gi|31241465 (18%)	NAD-dependent malate dehydrogenase	Similar to malate dehydrogenase mytochondrial in *D. melanogaster*
38.	gi|31204771 (4%)	S-adenosylhomocysteine hydrolase	Similar to CG9977-PA in *D. melanogaster*

Table 4. Continued

	Accession number[a), b)]	Domains/motifs	Features
39.	gi|31237882 (1%)	Threonyl-tRNA synthetase	Similar to threonyl-tRNA synthetase CG5353-PA in *D. melanogaster*
40.	gi|31212487 (3%)	Contains pyruvate kinase domain	Similar to pyruvate kinase CG7070-PA in *D. melanogaster*
41.	gi|31213491 (2%)	Contains staphylococcal nuclease and TUDOR domains	No significant similarity to other proteins
42.	gi|31213397 (4%)	Creatine kinases	Similar to arginine kinase CG32031-PC in *D. melanogaster*
43.	gi|31206811 (8%)	Fructose-biphosphate aldolase	Similar to aldolase CG6058-PF in *D. melanogaster*
44.	gi|31203043 (15%)	No conserved domains	Similar to salivary gland 1-like 4 protein
45.	**gi|31203045 (38%)**	No conserved domains	SAGLIN, contains a signal peptide
46.	**gi|31203049(45%)**	No conserved domains	Long form of the misannotated protein SG1-like 3 protein
47.	**gi|31234764 (30%)**	No conserved domains	Similar to 30-kDa protein
48.	gi|31214384 (6%)	No conserved domains	Orthologous to putative 53.7-kDa salivary protein in *A. stephensi*
49.	**gi|31203175 (11%)**	No conserved domains	Similar to gSG1b protein
50.	gi|21296299 (4%)	No conserved domains	No significant similarity to other proteins
51.	**gi|18873404 (25%)**	No conserved domains	Hypothetical protein (*A. gambiae*)

a) Proteins found both by in-gel and in-solution approaches are shown in bold

b) The percentage of sequence covered by identified peptides is indicated in parentheses

3 Results and discussion

3.1 Mass spectrometry-based characterization of salivary gland proteome

The most frequent strategy employed to study the molecules expressed in salivary glands of mosquitoes involves either random sequencing of clones from salivary gland cDNA libraries or use of a more specialized signal sequence trapping method for specific isolation of cDNAs encoding proteins with signal peptides [13–17]. A detailed review of high-throughput approaches to study salivary genes and proteins has been recently published [18]. In this study, we employed an MS-based approach to analyze the proteome of the female salivary gland. While both DNA and protein based methods would tend to identify the most abundant transcripts and proteins, it is possible that even when a transcript is identified, it might not be expressed at the protein level, as has been shown by the presence of a large number of transcribed pseudogenes in *C. elegans* [19]. However, if one is able to identify a protein, the corresponding genomic DNA can be automatically designated as a protein-coding region.

As illustrated in the schematic in Fig. 1, homogenized protein extracts from salivary glands from female *An. gambiae* mosquitoes were either digested directly in-solution with trypsin or first resolved by gel electrophoresis and subsequently digested with trypsin. In both cases, the complex peptide mixture was separated by LC and analyzed on a quadrupole TOF-MS. The data was searched against NCBI non-redundant database that contains known proteins as well as proteins that are encoded by predicted transcripts annotated by the EnsEMBL pipeline. Identification of proteins was achieved on the basis of at least one peptide with a MASCOT score above 30 or a manually validated mass spectrum, which could unambiguously provide a peptide sequence (Supplementary Table 1).

3.2 In-solution digestion approach for identification of proteins

In this approach, the salivary gland protein extract was digested in-solution prior to MS analysis as in 'shot-gun' proteomic approaches [20]. Following homogenization of salivary glands, the sample was digested with trypsin and then subjected to LC-MS/MS analysis. In this strategy, there is no prior fractionation of the sample by biochemical methods such as electrophoresis or chromatography that could theoretically lead to sample losses. Thus, an in-solution approach could be complimentary to in-gel digestion-based approaches.

Using the in-solution digestion strategy, we were able to identify nine known (Table 1) and 17 novel (Table 2) proteins. The large majority of these proteins either contained a pheromone/odorant-binding domain or lacked any conserved domains/motifs. In addition to identification of numerous D7-related proteins that have been well-described as salivary gland proteins, we found putative gVAG protein, histones H3, gSG proteins (gSG6, 7 and 1b), TRIO protein, sali-

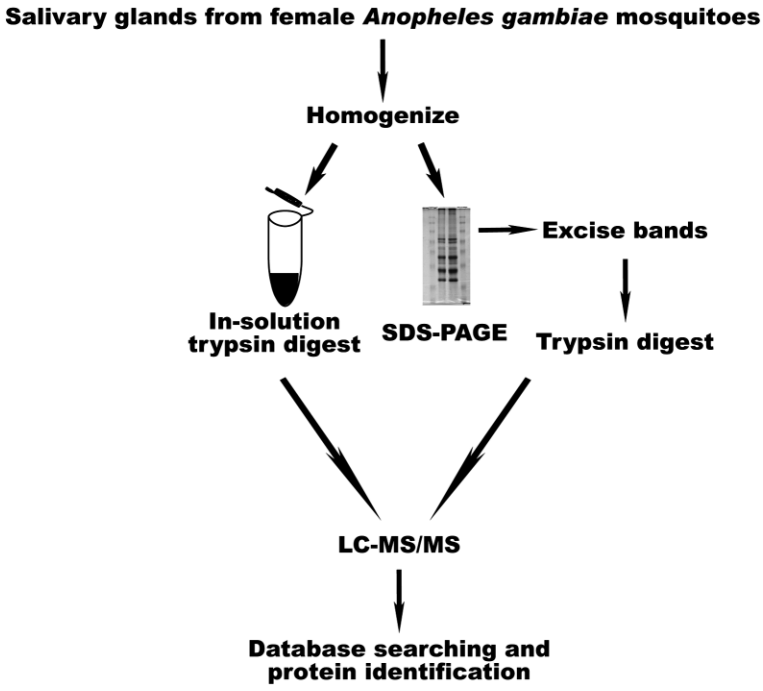

Figure 1. A schematic of proteomic characterization of the *An. gambiae* female mosquito salivary gland. After homogenization, the salivary gland protein extracts were resolved by SDS-PAGE as shown and digested with trypsin, or as an alternative strategy, digested in-solution by trypsin. The peptide mixture was analyzed by LC-MS/MS and the proteins identified by searching the MS spectra against the NCBInr protein database.

vary gland 1-like 3 protein, putative 5'-nucleotidase, protein disulfide isomerase, peroxidase and many others. In a recent study, the salivary gland proteome was investigated through an approach combining gel electrophoresis with Edman degradation [16]. We were able to identify all the proteins that were sequenced by the Edman degradation method using our in-solution approach (Table 1) or in-gel digestion approach (Table 3). In addition, we identified a number of other proteins that have not been previously described as components of salivary glands. Figure 2 illustrates the application of MS/MS for the identification of two salivary gland proteins. The highly complex peptide mixture derived from trypsin digestion of the tissue homogenate was separated using a RP column coupled online to the mass spectrometer as evidenced by the total ion chromatogram shown in Fig. 2A. The peptide (VGCSMWYWK) that eluted from the column at approximately 81 min was fragmented by CID and matched a protein designated putative gVAG protein precursor (gVAG precursor) (Fig. 2B). The MS/MS spectrum shown in Fig. 2C is derived from a peptide that eluted from the column around 97 min and was assigned as ANTFYTCFLGTSSLAGFK which matched D7-related 2 protein (D7-r2 protein). Ten tryptic peptides were found to correspond to the putative gVAG precursor (Supplementary Table 1), resulting in sequence coverage of 51%, whereas 15 tryptic peptides matched D-7 r2 protein (Supplementary Table 1) with 81% of sequence coverage, indicating that these are relatively abundant protein constituents of the salivary glands. One should bear in mind that putative gVAG precursor was identified in both in-solution and in-gel approaches, and that the Supplementary Table 1 lists the total number of peptides regardless of the method used.

Analysis of the putative gVAG protein sequence using the SMART program (http://smart.embl-heidelberg.de/) showed the presence of sperm-coating glycoprotein (SCP) family of extracellular domains that are widely found in eukaryotes including plants and yeast. In insects, this family is characterized as potent allergens that trigger allergic reactions to stings. The domain analysis as well as the presence of a signal peptide in this protein indicates that it is a secreted protein. Similarly, bioinformatics analysis of D-7 r2 protein shows that it contains the well-known pheromone/odorant-binding domain broadly distributed in several *Arthropod* species. As odorant molecules are primarily hydrophobic, pheromone/odorant binding proteins (OBP) greatly enhance the solubility of the odorants and facilitate its detection by the olfactory neurons membrane receptor [21, 22]. In insects, it is estimated that OBPs are found in the sesillum lymph, a fluid that bathes the chemosensory neurons, at relatively high concentrations in the range 10 mM [23]. This could account for the large number of peptides that match D-7 r2 protein in our database search results. The function of OBP is still not clear, although it could be related to the chemosensory detection of host preference by *Anopheles* mosquitoes.

The number of proteins identified in this study using the in-solution approach was somewhat lower than expected. We reason that this method allowed detection of only the most

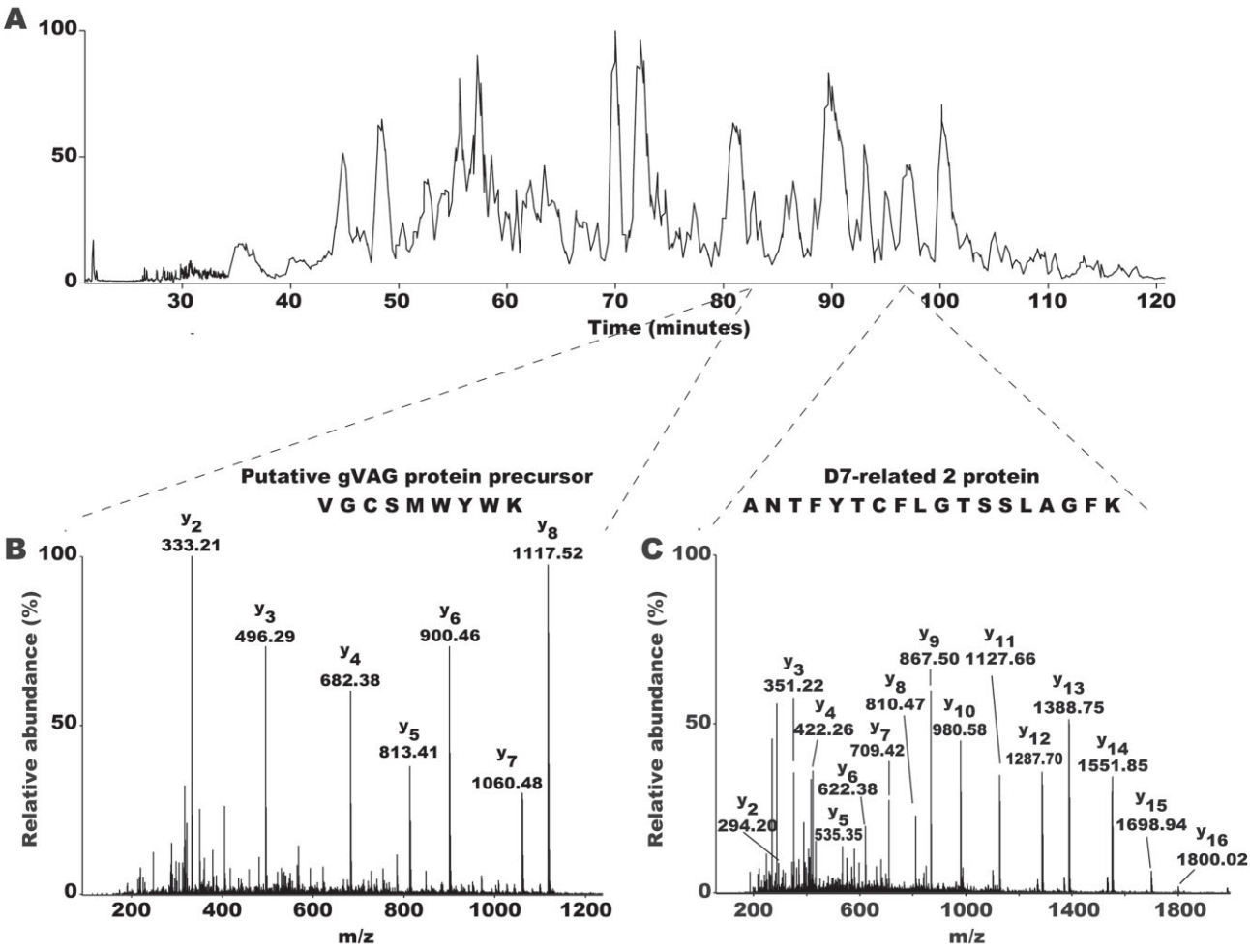

Figure 2. A gel-free approach for characterization of the salivary gland proteome. (A) Total ion chromatogram obtained from an LC-MS/MS run of the salivary gland homogenate followed by trypsin digestion in-solution. (B) Product ion MS/MS spectrum of a doubly charged ion at *m/z* 608.8 corresponds to the peptide sequence VGCSMWYWK, which matched a known protein designated as putative gVAG protein precursor. (C) MS/MS spectrum of a peak at *m/z* 994.0 corresponds to the peptide sequence ANTFYTCFLGTSSLAGFK. This sequence matched a known protein designated as D7-related 2 protein.

abundant proteins found in the salivary glands. The LC-MS/MS set-up used in our study consisted of a RP chromatographic column, which fractionated the peptides according to their hydrophobicity. Use of two independent stationary phases coupled together, in a so-called multidimensional protein identification method (MudPIT), could provide a larger number of protein identifications [24].

3.3 In-gel digestion approach for identification of proteins

In proteomics approaches, a primary strategy is the use of gel electrophoresis to resolve the protein sample according to the molecular mass and/or p*I* before the identification of the protein. One of the ways in which this can be achieved is by 2-DE. However, there are a number of limitations of 2-DE including the difficulty of visualizing very large or small proteins and very acidic or basic proteins [25]. Therefore, we decided to resolve the homogenate using SDS-PAGE. The extracts from salivary glands were separated by SDS-PAGE and the gel cut into 15 slices (Fig. 3A). Peptides derived from in-gel digestion of each slice were subsequently analyzed by LC-MS/MS in a manner identical to that described for in-solution digestion. Altogether, we were able to identify 56 proteins using this strategy, five of which were known proteins and 51 were novel proteins (Tables 3 and 4). Of these, 13 were previously identified using the in-solution method.

Figure 3B–D presents examples of MS data for identification of proteins. The MS/MS spectrum of the peak at *m/z* 469.74 is shown in Fig. 3B. The fragmentation pattern of this doubly charged ion provided the sequence of the peptide as FDWWER. This peptide matched a novel protein

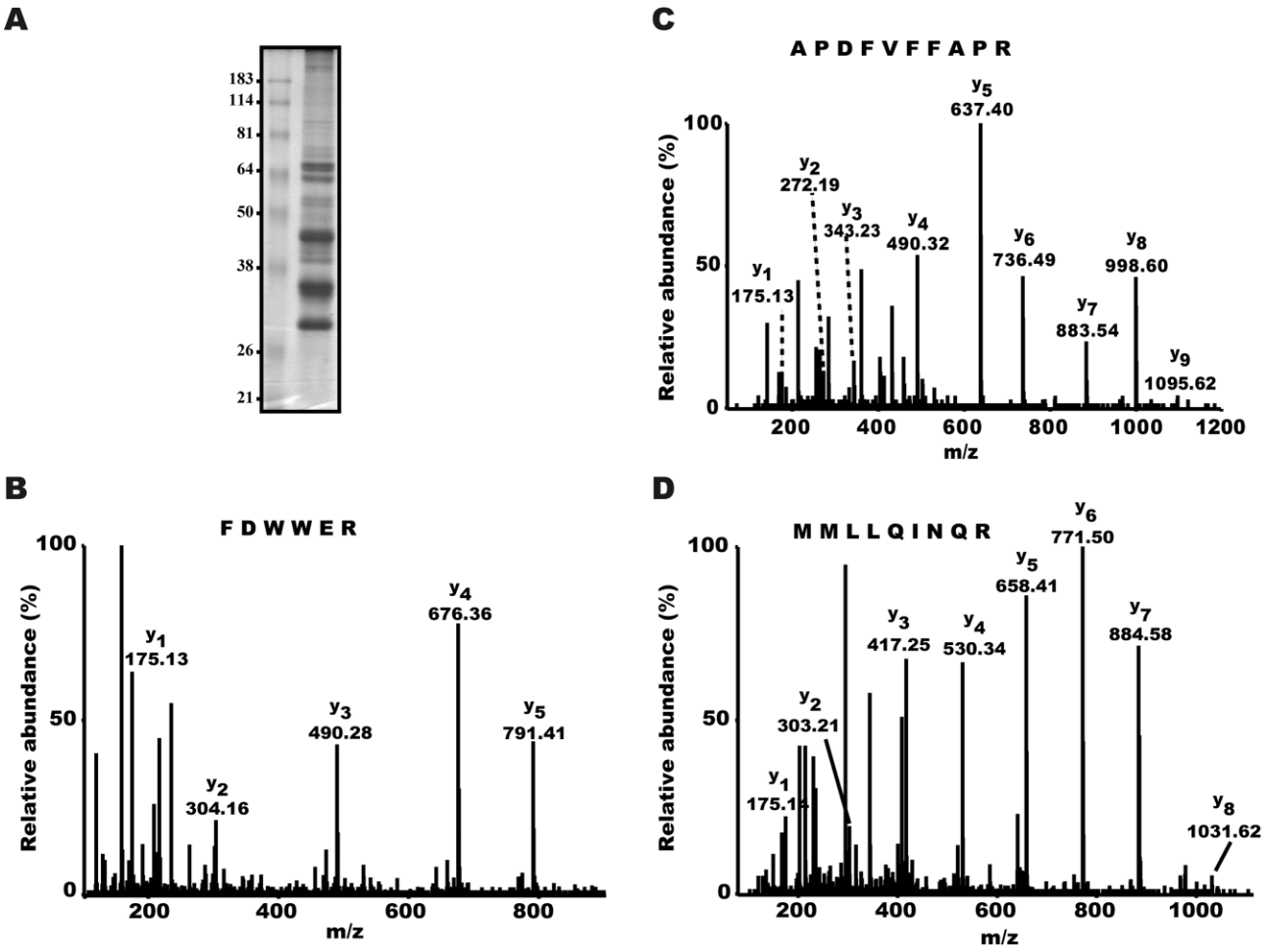

Figure 3. A gel-based approach for characterization of the salivary gland proteome. (A) SDS-PAGE gel of the salivary gland extract after staining with colloidal CBB is shown (right lane). The left lane shows markers whose molecular masses are indicated in kDa. (B) MS/MS spectrum of the peak at *m/z* 469.72 corresponds to the peptide sequence, FDWWER. This peptide matches a novel protein (gi|31241043), which is orthologous to a protein designated as probable maltase precursor in *Aedes aegypti*. (C) MS/MS spectrum of the doubly charged ion at *m/z* 583.83 corresponds to the peptide sequence APDFVFFAPR that matches a novel protein (gi|31203141) orthologous to a moesin in *D. melanogaster*. (D) MS/MS spectrum of the doubly charged ion at *m/z* 589.84 corresponds to the peptide sequence, MMLLQINQR, which matches a well-known salivary gland protein, SG1.

(gi|31241043) encoded by a predicted transcript in *An. gambiae*. This protein is 60% identical to a protein designated as probable maltase precursor (gi|126713) that was previously identified in *Aedes aegypti* [26, 27]. Likewise, five peptides also match this same entity (Supplementary Table 1). Another novel protein identified in this study was orthologous to a moesin in *Drosophila melanogaster*. The MS/MS spectrum clearly showed complete y fragment ions that readily assigns the amino acid sequence APDFVFFAPR (Fig. 3C), which matched this novel protein (gi|31203141). Although only this peptide was found to correspond to this protein (Supplementary Table 1), the unambiguous and clear MS/MS spectrum was enough to assure the identification. Figure 3D shows the MS/MS spectrum of one of the peptides (peak at *m/z* 589.84) whose sequence was determined to be MMLLQINQR. Searching this spectrum against NCBI nonredundant protein database identified a known salivary gland protein, SG1. Five other peptide sequences also matched this protein (Supplementary Table 1).

To characterize and achieve more sequence information on the proteins identified above a bioinformatics analysis was carried out. The novel protein ortholog to maltase precursor (gi|31241043) was first analyzed using SMART program, which predicted an alpha-amylase domain (Table 4). Further, a signal peptide was found from residues 1–28. The proteins that possess alpha-amylase domains are widely present in several invertebrate and vertebrate species. In mosquitoes, this type of protein is related to the sugar-meal digestion, and like other enzymes found in the salivary gland and midgut, its expression is regulated according to the

hematophagous mosquito feeding behavior [28, 29]. Similarly, we analyzed SG1 protein whose cDNA sequence was previously reported as salivary gland-specific and detected only in female mosquito [14, 30]. This protein is part of a protein family known as SG family that is found in other anophelines such as *An. stephensi* [16, 17]. SG1 protein sequence did not reveal any conserved domain and thus far no biological function has been ascribed to this protein. Software prediction found amino acid residues 1–16 as putative signal peptide indicative of secretion.

The other novel protein identified as an ortholog to a moesin in *D. melanogaster* contains a highly conserved ERM domain also known as ezrin/radixin/moesin protein domains. This domain is found in a number of cytoskeletal-associated proteins that associate with various proteins at the interface between the plasma membrane and the cytoskeleton. It has a conserved *N*-terminal domain involved in the linkage of cytoplasmic proteins to the membrane, whereas the *C*-terminal region has a sequence motif for actin binding. [31, 32]. It has been reported that Dmoesin, the ERM protein found in *D. melanogaster,* plays a crucial role in the actin organization during developmental process of oocytes, and mutations in this protein are lethal [33, 34]. In contrast to SG1 and the protein ortholog to maltase, we were unable to identify the presence of any signal peptide for this protein, suggesting that it is unlikely to be a secreted protein. This is in accordance with the cellular localization of ERM family of proteins, which are found in the nucleus and also at the cell-surface structures such as microvilli [32, 35].

MS/MS analysis identified a novel protein, SAGLIN, that was previously characterized by immunoaffinity purification using mAbs effective in blocking the infection of salivary glands by the sporozoite [36] (Okulate, M. *et al.* manuscript in preparation). In an *in vivo* bioassay, the mAb raised against the 100 kDa protein inhibited *Plasmodium yoelii* sporozoite invasion of salivary glands by 73%. These results show that *An. gambiae* salivary gland proteins are accessible to mAbs that inhibit sporozoite invasion of the salivary glands, and suggest alternate targets for blocking the transmission of malaria by this most competent malaria vector [36].

3.4 Validation of predicted transcripts in the *An. gambiae* using MS-derived data

Currently, 64 known salivary gland proteins are deposited in the Swiss-Prot database and 16 424 transcripts are available in the TrEMBL collection. In this work, we present direct evidence for the presence of 14 known and 68 novel proteins in the salivary gland. Protein sequence coverage observed for the identified proteins varied from 1% to 81% (Tables 1–4). A validation of other predicted transcripts could similarly be accomplished through the use of direct peptide sequence data such as that obtained by MS/MS in our study.

Overall, searching of the mass spectrometry data against protein database made possible the identification of 13 proteins solely from the gel-free approach, 43 proteins solely after SDS-PAGE and 13 proteins from both approaches (Fig. 4A). The majority of the identified proteins present at least two peptide sequences and 19 out of 69 identified entities presented only one unique peptide (Supplementary Table 1). In the gel-free approach, the peptide sequences matched nine known proteins (Table 1) and 17 predicted transcripts in *An. gambiae* (Table 2); likewise, in the approach using SDS-PAGE, the peptide sequences matched five known proteins (Table 3) and 51 predicted transcripts in *An. gambiae* (Table 4). These data indeed confirm that these two methods are complementary as the use of one strategy alone could lead to loss of some proteins identified by the other strategy.

3.5 Comparative genomic analysis of *An. gambiae* salivary gland proteome

We have performed a comparative genomic analysis of *An. gambiae* and *An. stephensi* transcriptome obtained from salivary gland cDNA sequences and find that 48% of the proteins had easily identifiable homologs in both species. The degree of homology ranged from 42% to 92%, indicating that most of these proteins are highly conserved in these two hematophagous mosquito species. Previous comparison of the degree of identity of housekeeping and salivary gland gene products between *An. stephensi* and *An. gambiae* [17] indicates that the former one presented 93.11 $\pm$ 5.93% identity in average, whereas the salivary gland genes presented 62.4 $\pm$ 15.4% identity. We were unable to find orthologs of approximately 12% of the identified proteins including gi|31228364, gi|21292024, gi|31209765, gi|31241427, gi|31214711, gi|31213491, gi|31203045 and gi|21296299 that were present in *An. gambiae* (Table 4). This implies either that these proteins are specific to *An. gambiae* or, more likely, are not identifiable because the genome sequence of *An. stephensi* has not yet been completed. Among several haematophagous insect species that promote transmission of disease to humans, *An. gambiae* is the only species whose genome is completely sequenced. Sequencing of ESTs and genomic contigs from the yellow fever transmitting mosquito, *Aedes aegypti*, is an ongoing project (http://www.nd.edu/~dseverso/genome.html) that should provide additional comparative information about the proteomes of these important mosquito vectors.

3.6 Functional assignments for the salivary gland proteins

An. gambiae is a dipteran hematophage and its saliva is rich in anti-inflammatory and anti-hemostatic enzymes which interfere with blood coagulation and inhibit the pain response, allowing the mosquito to have a blood-meal with a minimal chance of detection [4, 37]. Proteases in the midgut, orthologs of trypsin and chymotrypsin found in higher organisms, are involved in rapid digestion [38], whereas others are involved in immune response [39]. The salivary gland and midgut regions of the mosquito have attracted a great deal of attention, as this is where the *Plasmodium* develops

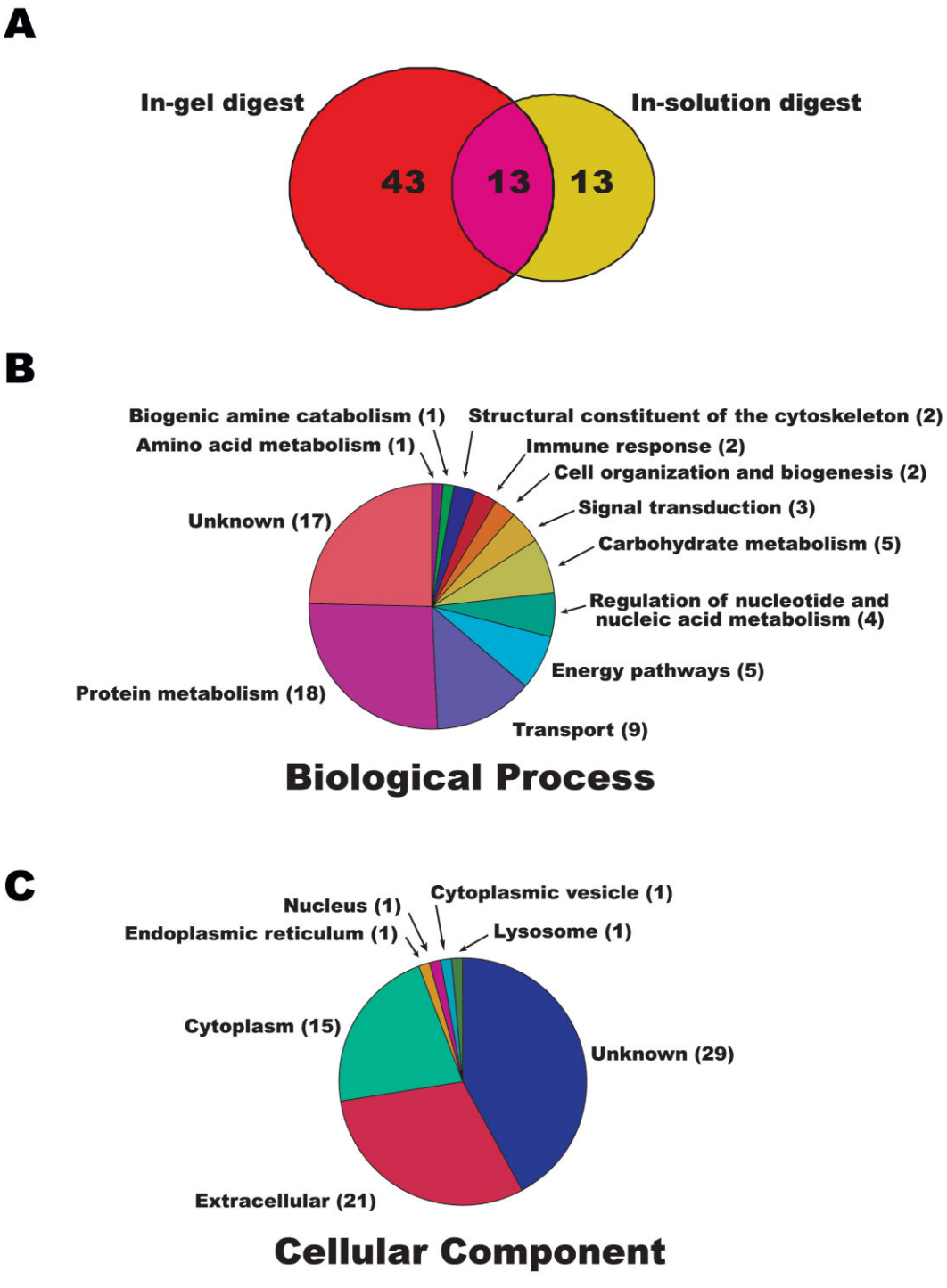

Figure 4. A comparison of in-solution and in-gel digestion approaches and functional annotation of the salivary gland proteome. A Venn diagram with the number of proteins identified using the in-solution and in-gel digestion strategy (A). Distribution of the identified proteins from the salivary gland grouped according to biological process (B) and cellular component (C). The number of identified proteins is in parentheses.

and matures in its vector. After a period of extensive development in the midgut, the sporozoites migrate to the salivary gland likely using species-specific receptor-mediated interactions for recognition and invasion. They are then injected into the bloodstream of the host when the mosquito takes its next blood meal. Various groups have worked on the salivary

gland transcriptome of different *Anopheles* species [16, 17, 27, 40]. These studies have led to characterization of three types of gene products: secretory molecules, housekeeping gene products and proteins of unknown function. Secreted proteins include amylase, calreticulin, selenoprotein, mucin-like protein, 30 kDa allergen, antigen 5, D7-related proteins, SG-like proteins and putative secretory proteins. Housekeeping gene products included thioredoxin, tetraspanin, hemopexin, heat shock proteins, TRIO and MBF proteins. Proteins of unknown function are those that are usually not similar to any known protein and generally have no obvious protein domain or motif that can provide some clues regarding protein function.

We provide a functional annotation of the *An. gambiae* genome based on the salivary gland proteins identified in this study. Gene ontology is now widely used to describe protein function in a standardized format (http://www.geneontology.org/) [41]. Thus, we performed a bioinformatics analysis to assign a biological process to as many proteins as possible (Fig. 4B). A large proportion of the identified proteins were involved in protein, carbohydrate and nucleic acid metabolism, transport or energy pathways. Almost 25% of the proteins could be not be ascribed any biological function. We also assigned a cellular component (*i.e.*, subcellular localization) to each protein either based on the literature or the presence of particular domains/motifs (Fig. 4C). As expected, the majority of the proteins were classified as extracellular proteins (Fig. 4C). D7r family proteins, apyrases and proteins of the salivary gland-like (SG-like) family were the commonest extracellular proteins. Proteins involved in translation and protein folding were the predominant cytoplasmic proteins with a small number of proteins classified as nuclear, vesicular or lysosomal proteins. Approximately 40% of proteins could not be assigned any specific localization because of lack of any distinctive features and lack of homology to other known proteins.

4 Concluding remarks

In the post-genomic era, MS has emerged as a powerful tool for high-throughput analyses of proteomes [42, 43]. In this report, we describe the first such MS-based strategy for characterizing the proteome of salivary gland of *An. gambiae*, the major malaria vector in sub-Saharan Africa. Using two complementary strategies for protein preparation, we were able to identify a number of known proteins. Interestingly, the vast majority of identified proteins was novel, being represented only as predicted transcripts in the databases. A comparison of our dataset with a cDNA based strategy [16] revealed that 41 of the proteins reported in this study were not identified previously as cDNAs. This illustrates the complementary nature of different techniques for identification of genes and gene products. Gene ontology assignments of the identified proteins demonstrated that the majority of the identified proteins are likely to be involved in transport or metabolism and are located extracellularly or are cytosolic. We have demonstrated that MS provides valuable data that can be used to validate genome annotations and to discover novel proteins in a high-throughput manner. Novel proteins identified from such approaches can subsequently be tested in strategies developed to control pathogen transmission and disease.

A.P. and N.K. were supported by a pilot project grant from the Johns Hopkins Malaria Research Institute. N.K. was also supported by a National Institutes of Health grant. We thank Suraj Peri for the initial assistance on genome analysis and Dr. Jakob Bunkenborg for providing a script to combine mass spectra before database searching. We are grateful to Sun Microsystems for providing us a computer cluster under the Academic Equipment Grant mechanism. We are indebted to John Kloss for setting up the cluster and maintenance of MASCOT program. We thank everyone in the Pandey lab for fruitful discussions.

References

[1] *Wkly Epidemiol. Rec.* 1994, *69*, 309–314.

[2] Morel, C. M., Toure, Y. T., Dobrokhotov, B., Oduola, A. M., *Science* 2002, *298*, 79.

[3] Howard, D. H., Scott, R. D., 2nd, Packard, R., Jones, D., *Clin. Infect. Dis.* 2003, *36*, S4–10.

[4] Budiansky, S., *Science* 2002, *298*, 80–86.

[5] Rosenberg, R., *Am. J. Trop. Med. Hyg.* 1985, *34*, 687–691.

[6] Sidjanski, S. P., Vanderberg, J. P., Sinnis, P., *Mol. Biochem. Parasitol.* 1997, *90*, 33–41.

[7] Clements, A. N., *Biology of Mosquitoes: Development, Nutrition and Reproduction*, Chapman and Hall, London 1992, p. *536*.

[8] Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G. *et al.*, *Science* 2002, *298*, 129–149.

[9] Shevchenko, A., Wilm, M., Vorm, O., Mann, M., *Anal. Chem.* 1996, *68*, 850–858.

[10] Carr, S., Aebersold, R., Baldwin, M., Burlingame, A. *et al.*, *Mol. Cell. Proteomics* 2004, *3*, 531–533.

[11] Perkins, D. N., Pappin, D. J., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551–3567.

[12] Schultz, J., Milpetz, F., Bork, P., Ponting, C. P., *Proc. Natl. Acad. Sci. USA* 1998, *95*, 5857–5864.

[13] Calvo, E., Andersen, J., Francischetti, I. M., de Lara Capurro, M. *et al.*, *Insect Mol. Biol.* 2004, *13*, 73–88.

[14] Arca, B., Lombardo, F., de Lara Capurro, M., della Torre, A. *et al.*, *Proc. Natl. Acad. Sci. USA* 1999, *96*, 1516–1521.

[15] Arca, B., Lombardo, F., Lanfrancotti, A., Spanos, L. *et al.*, *Insect Mol. Biol.* 2002, *11*, 47–55.

[16] Francischetti, I. M., Valenzuela, J. G., Pham, V. M., Garfield, M. K. *et al.*, *J. Exp. Biol.* 2002, *205*, 2429–2451.

[17] Valenzuela, J. G., Francischetti, I. M., Pham, V. M., Garfield, M. K. *et al.*, *Insect Biochem. Mol. Biol.* 2003, *33*, 717–732.

[18] Valenzuela, J. G., *Insect Biochem. Mol. Biol.* 2002, *32*, 1199–1209.

[19] Mounsey, A., Bauer, P., Hope, I. A., *Genome Res.* 2002, *12*, 770–775.

[20] McDonald, W. H., and Yates, J. R., 3rd, *Curr. Opin. Mol. Ther.* 2003, *5*, 302–309.

[21] Zhang, S., Maida, R., Steinbrecht, R. A., *Chem. Senses* 2001, *26*, 885–896.

[22] Vogt, R. G., *J. Chem. Ecol.* 2002, *28*, 2371–2376.

[23] Xu, P. X., Zwiebel, L. J., Smith, D. P., *Insect Mol. Biol.* 2003, *12*, 549–560.

[24] Wolters, D. A., Washburn, M. P., Yates, J. R., 3rd, *Anal. Chem.* 2001, *73*, 5683–5690.

[25] Ong, S. E., Pandey, A., *Biomol. Eng.* 2001, *18*, 195–205.

[26] Grossman, G. L., James, A. A., *Insect Mol. Biol.* 1993, *1*, 223–232.

[27] Valenzuela, J. G., Pham, V. M., Garfield, M. K., Francischetti, I. M. *et al.*, *Insect Biochem. Mol. Biol.* 2002, *32*, 1101–1122.

[28] Ribeiro, J. M., *Insect Biochem. Mol. Biol.* 2003, *33*, 865–882.

[29] Shen, Z., Edwards, M. J., Jacobs-Lorena, M., *Insect Mol. Biol.* 2000, *9*, 223–229.

[30] Lanfrancotti, A., Lombardo, F., Santolamazza, F., Veneri, M. *et al.*, *FEBS Lett.* 2002, *517*, 67–71.

[31] Bretscher, A., Chambers, D., Nguyen, R., Reczek, D., *Annu. Rev. Cell. Dev. Biol.* 2000, *16*, 113–143.

[32] Tsukita, S., Yonemura, S., *J. Biol. Chem.* 1999, *274*, 34507–34510.

[33] Miller, K. G., *Trends Cell Biol.* 2003, *13*, 165–168.

[34] Polesello, C., Payre, F., *Trends Cell Biol.* 2004, *14*, 294–302.

[35] Batchelor, C. L., Woodward, A. M., Crouch, D. H., *Exp. Cell Res.* 2004, *296*, 208–222.

[36] Brennan, J. D., Kent, M., Dhar, R., Fujioka, H. *et al.*, *Proc. Natl. Acad. Sci. USA* 2000, *97*, 13859–13864.

[37] Ribeiro, J. M., and Francischetti, I. M., *Annu. Rev. Entomol.* 2003, *48*, 73–88.

[38] Vizioli, J., Catteruccia, F., della Torre, A., Reckmann, I. *et al.*, *Eur. J. Biochem.* 2001, *268*, 4027–4035.

[39] Gorman, M. J., Paskewitz, S. M., *Insect Biochem. Mol. Biol.* 2001, *31*, 257–262.

[40] Calvo, E., Andersen, J., Francischetti, I. M., De, L. C. M. *et al.*, *Insect Mol. Biol.* 2004, *13*, 73–88.

[41] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D. *et al.*, *Nat. Genet.* 2000, *25*, 25–29.

[42] Mann, M., Hendrickson, R. C., Pandey, A., *Annu. Rev. Biochem.* 2001, *70*, 437–473.

[43] Zhang, H., Yan, W., Aebersold, R., *Curr. Opin. Chem. Biol.* 2004, *8*, 66–75.

