An Insight into the Sialotranscriptome of *Simulium nigrimanum,* a Black Fly Associated with Fogo Selvagem in South America

José M. C. Ribeiro,* Jesus G. Valenzuela, Van M. Pham, Lindsay Kleeman, Kent D. Barbian, Amanda J. Favreau,
Donald P. Eaton, Valeria Aoki, Gunter Hans-Filho, Evandro A. Rivitti, and Luis A. Diaz

*Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, Rockville, Maryland;
Genomics Unit, Research Technologies Section, Rocky Mountain Laboratories, Hamilton, Montana; Departments of Dermatology,
University of North Carolina at Chapel Hill, North Carolina; Department of Dermatology, Universidade de São Paulo, Brazil;
Universidade Federal de Mato Grosso do Sul and the Wildlife Conservation Society of Brazil, Campo Grande, MS, Brazil*

Abstract. Pemphigus foliaceus is a life threatening skin disease that is associated with autoimmunity to desmoglein, a skin protein involved in the adhesion of keratinocytes. This disease is endemic in certain areas of South America, suggesting the mediation of environmental factors triggering autoimmunity. Among the possible environmental factors, exposure to bites of black flies, in particular *Simulium nigrimanum* has been suggested. In this work, we describe the sialotranscriptome of adult female *S. nigrimanum* flies. It reveals the complexity of the salivary potion of this insect, comprised by over 70 distinct genes within over 30 protein families, including several novel families, even when compared with the previously described sialotranscriptome of the autogenous black fly, *S. vittatum*. The uncovering of this sialotranscriptome provides a platform for testing pemphigus patient sera against recombinant salivary proteins from *S. nigrimanum* and for the discovery of novel pharmacologically active compounds.

INTRODUCTION

Pemphigus foliaceus, also known by the Portuguese name "fogo selvagem," is a life threatening autoimmune disease leading to keratinocyte detachment and skin blistering.[1] Patients have increased antibody against desmoglein, a structural protein important in intercellular adhesion.[2,3] This is a rare disease in general, but endemic in some Amerindian reservations in Brazil. The patchy spatial distribution of this condition in genetically related groups indicated an environmental component causing the disease.[4] Among the environmental variables, exposure to black fly bites, in particular to the anthropophilic species *Simulium nigrimanum*, was shown to correlate with anti-desmoglein titers in an epidemiological study.[5]

Saliva of blood-sucking arthropods contain a vast array of pharmacologically active compounds that disarms their hosts hemostasis, the physiological reaction preventing blood loss after tissue injury, and inflammation, which can lead to enhanced hemostasis and host defense reactions.[6,7] The advent of transcriptome methods has led to characterization of the complexity of the salivary potion of blood feeding arthropods, which number several dozen different proteins for sand flies, near 100 for mosquitoes and black flies, and hundreds in the case of hard ticks.[8] This salivary potion contains enzymes such as apyrase, which degrades adenosine diphosphate (ADP) and adenosine triphosphate (ATP) (powerful agonists of platelet and neutrophil aggregation), hyaluronidase (which facilitates the spreading of the salivary pharmaceuticals into the feeding pool), vasodilators in various forms, different platelet aggregation antagonists, complement activation inhibitors, and agonist chelators, or kratagonists, proteins that bind different hemostasis and inflammation agonists such as biogenic

amines (serotonin, norepinephrine, and histamine), inflammatory leukotrienes, and thromboxane A$_2$, a potent platelet agonist.[8]

Comparative transcriptome analyses of different mosquito species and genera indicate that the evolution of salivary proteins has occurred at a very fast pace, possibly caused by the immune pressure of their hosts. Indeed, comparisons between *Anopheles*, *Culex*, and *Aedes* showed that each contained genus-specific protein families, and even subgenus-specific families.[9–14] Regarding black flies, a single sialotranscriptome (from the Greek sialo = saliva) has been analyzed so far, from the North American species *Simulium vittatum*.[15] It is the goal of this work to describe the sialotranscriptome of *S. nigrimanum*, to compare it with that of *S. vittatum*, and to provide candidate proteins that might be the trigger of anti-desmoglein antibodies in fogo selvagem.

MATERIALS AND METHODS

Chemicals. Standard laboratory chemicals were purchased from Sigma Chemicals (St. Louis, MO) if not specified otherwise.

Black flies. Adult *S. nigrimanum* (113 flies) were collected near streams and homes in the Terena Ameridian community of Aldeia Limão Verde, municipality of Aquidauana, Mato Grosso do Sul state, Brazil. From October 6 to 8, 2007, flies were captured with aspirators from the exposed limbs of coauthor, DPE. The salivary glands were dissected in phosphate-buffered saline (PBS), immediately transferred to 50 μL of RNAlater (Ambio, Inc., Austin, TX) and kept refrigerated for 7 days. After transport to the United States, the glands were frozen and shipped to the Laboratory of Malaria and Vector Research, Rockville, MD.

Library construction. Salivary gland RNA, extracted from 193 intact glands, was isolated using the Micro-FastTrack mRNA isolation kit (Invitrogen, San Diego, CA). The polymerase chain reaction (PCR)-based cDNA library was made following the instructions for the SMART cDNA library construction kit (Clontech, Palo Alto, CA). This system uses oligoribonucleotide (SMART IV) to attach an identical sequence at the 5′ end of each reverse-transcribed cDNA

*Address correspondence to José M. C. Ribeiro, 12735 Twinbrook Parkway room 2E32D, Rocvkville, MD 20852. E-mail: jribeiro@niaid.nih.gov
This paper should be a guide to browsing the supplemental files S1 and S2, found at: http://exon.niaid.nih.gov/transcriptome/S_nigrimanum/S1/Sn-S1-web.xls; http://exon.niaid.nih.gov/transcriptome/S_nigrimanum/S2/Sn-S2-web.xls.

strand. This sequence is then used in subsequent PCR reactions and restriction digests.

First-strand synthesis was carried out using PowerScript reverse transcriptase at 42°C for 1 hour in the presence of the SMART IV and CDS III (3′) primers. Second-strand synthesis was performed using a long distance (LD) PCR-based protocol, using Advantage Taq polymerase (Clontech) mix in the presence of the 5′ PCR primer and the CDS III (3′) primer. The cDNA synthesis procedure resulted in creation of *SfiI A* and *B* restriction enzyme sites at the ends of the PCR products that are used for cloning into the phage vector (lambda TriplEx2 vector, Clontech). The PCR conditions were as follows: 95°C for 1 min; 24 cycles of 95°C for 10 sec, 68°C for 6 min. A small portion of the cDNA obtained by PCR was analyzed on a 1.1% agarose gel to check quality and range of cDNA synthesized. Double-stranded cDNA was immediately treated with proteinase K (0.8 µg/mL) at 45°C for 20 min, and the enzyme was removed by ultrafiltration though a Microcon (Amicon Inc., Beverly, CA) YM-100 centrifugal filter device. The cleaned, double-stranded cDNA was then digested with *SfiI* at 50°C for 2 hours, followed by size fractionation on a ChromaSpin-400 column (Clontech) into small (S), medium (M), and large (L) transcripts based on their electrophoresis profile on a 1.1% agarose gel. Selected fractions were pooled and concentrated using a Microcon YM-100.

The concentrated cDNA mixture was ligated into the λ TriplEx2 vector (Clontech), and the resulting ligation mixture was packaged using the GigaPack III Plus packaging extract (Stratagene, La Jolla, CA) according to the manufacturer's instructions. The packaged library was plated by infecting log-phase XL1- Blue *Escherichia coli* cells (Clontech). The percentage of recombinant clones was determined by blue-white selection screening on LB/MgSO$_4$ plates containing X-gal/IPTG. Recombinants were also determined by PCR, using vector primers, PT2F1 (AAG TAC TCT AGC AAT TGT GAG C) and PT2R1 (CTC TTC GCT ATT ACG CCA GCT G flanking the inserted cDNA, with subsequent visualization of the products on a 1.1% agarose/EtBr gel.

cDNA sequencing. Twenty-four 96-well plates were prepared for cyclo sequencing, each containing 94 clones and two DNA controls, as follows: The cDNA library was plated on LB/MgSO4 plates containing X gal/IPTG to an average of 250 plaques per 150 mm Petri plate. Recombinant (white) plaques were randomly selected and transferred to 96-well microtiter plate (Nunc, Rochester, NY) containing 75 µL of ultrapure water (KD Medical, Columbia, MD) per well. The plates were covered and placed on a gyrating shaker for 30 min at room temperature. The phage suspension was either immediately used for PCR or stored at 4°C for future use.

To amplify the cDNA using a PCR reaction, 4 µL of the phage sample was used as a template. The primers were sequences from the λ TriplEx2 vector and named PT2F1 (AAG TAC TCT AGC AAT TGT GAG C) and PT2R1 (CTC TTC GCT ATT ACG CCA GCT G), positioned at the 5′ end and the 3′ end of the cDNA insert, respectively. The reaction was carried out in a 96-well PCR microtiter plate (Applied Biosystems, Inc., Foster City, CA) using FastStart Taq polymerase (Roche Diagnostics) on a GeneAmp PCR system 9700 (Perkin Elmer Corp., Foster City, CA). The PCR conditions were 1 hold of 75°C for 3 min; 1 hold 94°C for 4 min, 33 cycles of 94°C for 1 min, 49°C for 1 min; 72°C for 2 min. The amplified products were analyzed on a 1.5% agarose/EtBr gel. Clones were PCR

amplified, and the ones showing single band were selected for sequencing. Approximately 200–250 ng of each PCR product was transferred to a 96-well PCR microtiter plate (Applied Biosystems) and frozen at −20°C. Samples were shipped on dry ice to the Rocky Mountain Laboratories Genomics Unit with primer (PT2F3, TCT CGG GAA GCG CGC CAT TGT) and template combined together in an ABI 96-well Optical Reaction Plate (P/N 4306737) following the manufacturer's recommended concentrations. Sequencing reactions were set up as recommended by Applied Biosystems BigDye Terminator v3.1 Cycle Sequencing Kit by adding 1 µL ABI BigDye Terminator Ready Reaction Mix v3.1 (P/N 4336921), 1.5 µL 5× ABI Sequencing Buffer (P/N 4336699), and 3.5 µL of water for a final volume of 10 µL. Cycle sequencing was performed at 96°C for 10 sec, 50°C for 5 sec, 60°C for 4 min for 27 cycles on either a Bio-Rad Tetrad 2 (Bio-Rad Laboratories, Hercules, CA) or ABI 9700 (Applied Biosystems) thermal cycler. Fluorescently-labeled extension products were purified following Applied Biosystems BigDye XTerminator Purification protocol and subsequently processed on an ABI 3730xL DNA Analyzer (Applied Biosystems). The AB1 file generated for each sample from the 3730xL DNA Analyzer was provided to researchers in Rockville, MD through a secure network drive for all subsequent downstream sequencing analysis. In addition to the sequencing of the cDNA clones, primer extension experiments were performed in selected clones to further extend sequence coverage.

Bioinformatic tools and procedures. Expressed sequence tags (EST) were trimmed of primer and vector sequences. The BLAST tool,[16] CAP3 assembler,[17] and ClustalW[18] software were used to compare, assemble, and align sequences, respectively. Phylogenetic analysis and statistical neighbor-joining (NJ) bootstrap tests of the phylogenies were done with the Mega package.[19] For functional annotation of the transcripts we used the tool BlastX[16] to compare the nucleotide sequences to the non-redundant (NR) protein database of the National Center for Biotechnology Information (NCBI, National Library of Medicine, NIH,) and to the Gene Ontology (GO) database.[20] The tool, reverse position specific Blast (RPS-BLAST)[16] was used to search for conserved protein domains in the Pfam,[21] SMART,[22] Kog,[23] and conserved domains databases (CDD).[24] We have also compared the transcripts with other subsets of mitochondrial and rRNA nucleotide sequences downloaded from NCBI. Segments of the three-frame translations of the EST (because the libraries were unidirectional, 6-frame translations were not used), starting with a methionine found in the first 300 predicted amino acids (AAs), or the predicted protein translation in the case of complete coding sequences, were submitted to the SignalP server[25] to help identify translation products that could be secreted. O-glycosylation sites on the proteins were predicted with the program NetOGlyc.[26] Functional annotation of the transcripts was based on all the comparisons above. Following inspection of all these results, transcripts were classified as either Secretory (S), Housekeeping (H), or of Unknown (U) function, with further subdivisions based on function and/or protein families. Codon volatility was calculated as previously described.[27]

RESULTS AND DISCUSSION

cDNA library characteristics. A total of 2,039 clones out of 2,350 that were sequenced yielded good quality sequences and

were used to assemble a database (Supplemental Table S1) that yielded 827 clusters of related sequences, 651 of which contained only one EST. The consensus sequence of each cluster is named either a contig (deriving from two or more sequences) or a singleton (deriving from a single sequence). For sake of simplicity, this work uses "cluster" or "contig" to denote sequences derived from both consensus sequences and singletons. The 827 clusters were compared using the program BlastX, BlastN, or RPS-BLAST[16] to the non-redundant protein database of the NCBI (NR), a gene ontology database (GO),[20] the conserved domains database of the NCBI (CDD),[24] and a custom-prepared subset of the NCBI nucleotide database containing either mitochondrial or rRNA sequences.

Because the libraries used are unidirectional, three-frame translations of the dataset were also derived, and open reading frames (ORFs) starting with a methionine and longer than 40 AAs residues were submitted to SignalP server[25] to help identify putative-secreted proteins. The EST assembly, BLAST, and signal peptide results were loaded into an Excel (Microsoft, Redman, OR) spreadsheet for manual annotation and are provided in Supplemental Table S1.

Four categories of expressed genes derived from the manual annotation of the contigs were created (Table 1 and Figure 1). The putatively secreted (S) category contained 29.6% of the clusters and 61.5% of the sequences, with an average number of 5.1 sequences per cluster. The housekeeping (H) category had 27.9% and 20.2% of the clusters and sequences, respectively, and an average of 1.8 sequences per cluster. Four singletons were classified as transposable element/Viral (TEV) products, constituting less than 0.5% of the ESTs or contigs. Transposable elements (TE) have been a common finding in sialotranscriptomes and most probably reflect regulatory transcripts repressing transposition rather than active transposition.[28] For example, the transcript sn_contig_826 matches a Gypsy transposon polymerase, but has a stop codon. More specific to *Simulium* is the transcript sn-contig_406 that matches polyproteins of picorna-like virus specific to insects.[29] The picornaviridae is a ancient family affecting prokaryotes and eukaryotes, both animal and plants[30] and are responsible for varied diseases such as hepatitis (Hepatitis A) and poliomyelitis[31] in humans and paralytic diseases in insects.[29,32] Finally, 42% of the clusters, containing 18% of all sequences, were classified as unknown (U), because no functional assignment could be made. This category had an average of 1.1 sequences per cluster. A good proportion of these transcripts could derive from 3′ or 5′ untranslated regions of genes of the previous two categories, as was recently indicated for a sialotranscriptome of *Anopholes gambiae*.[12]

Housekeeping (H) genes. The 231 clusters (comprising 411 ESTs) attributed to H genes expressed in the salivary glands of *S. nigrimanum* were further divided into 18 subgroups according to function (Table 2). Not surprisingly for an organ
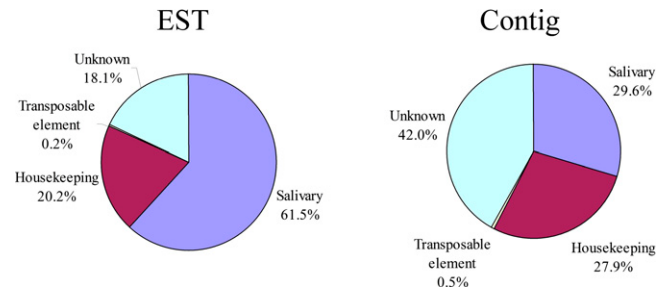

FIGURE 1. Functional class distribution of expressed sequence tags (EST) or assembled contigs (Contigs) deriving from a salivary gland cDNA library from adult *Simulium nigrimanum* black flies. This figure appears in color at www.ajtmh.org.

specialized for the secretion of polypeptides, the two larger sets were associated with protein synthesis machinery (49 clusters containing 117 ESTs) and energy metabolism (43 clusters containing 80 ESTs), a pattern also observed in other sialotranscriptomes.[13,33,34] We have arbitrarily included a group of 111 ESTs (41clusters) in the H category that represent highly conserved proteins of unknown function, presumably associated with cellular function. They are named conserved proteins of unknown function in Supplemental Table S1, immediately preceding the clusters of the TEV class. These sets may help functional identification of the "conserved hypothetical" proteins as previously reviewed by Galperin and Koonin.[35] The complete list of all 231 gene clusters, along with further information about each, is given in Supplemental Table S1.

Possibly secreted (S) class of expressed genes. Inspection of Supplemental Table S1 indicates the expression of several expanded gene families, including those coding for Kunitz-domain containing polypeptides, antigen-5 family members, odorant-binding/D7 protein families, vasodilatory proteins of the *Simulium vittatum* erythema protein (SVEP) family, low complexity protein families, and several protein families previously found only in *S. vittatum*[15] (Table 3).

Detailed analysis of the *S. nigrimanum* sialome. Several clusters of sequences coding for housekeeping and putative

TABLE 1

Functional classification of transcripts originating from the salivary glands of *Simulium nigrimanum*

Class	Number of contigs	Number of ESTs	EST's/contig
Salivary	245	1255	5.1
Housekeeping	231	411	1.8
Transposable element/virus	4	4	1.0
Unknown	347	369	1.1

TABLE 2

Functional classification of Housekeeping transcripts originating from the salivary glands of *Simulium nigrimanum*

Class	Number of contigs	Number of ESTs	EST's/contig
Protein synthesis machinery	49	117	2.4
Unknown, conserved	41	111	2.7
Metabolism, energy	43	80	1.9
Signal transduction	23	25	1.1
Protein modification machinery	14	14	1.0
Protein export machinery	12	12	1.0
Transcription factor	9	11	1.2
Nuclear regulation	6	6	1.0
Proteasome machinery	5	6	1.2
Transporters/storage	6	6	1.0
Cytoskeletal	6	6	1.0
Transcription machinery	4	4	1.0
Metabolism, carbohydrate	3	3	1.0
Metabolism, nucleotide	3	3	1.0
Oxidant metabolism/detoxification	2	2	1.0
Metabolism, amino acid	2	2	1.0
Metabolism, lipid	2	2	1.0
Metabolism, intermediate	1	1	1.0
Total	231	411	

TABLE 3
Functional classification of putative secreted transcripts originating from the salivary glands of *Simulium nigrimanum*

Class	Number of ESTs
Subclass	
Protease inhibitor domains	
Kunitz domains	91
Enzymes	
Serine proteases	82
Endonuclease	1
Destabilase	5
5′-nucleotidase/apyrase	19
Amylase	52
Antigen 5 family	31
Lipocalin	1
Immunity related	
Pattern recognition	2
Antimicrobial	36
Families unique to blood sucking Nematocera	
D7/OBP	317
Aegyptin family	43
Families unique to black flies	
SVEP*	86
Other *Simulium*-specific proteins	171
Families with low complexity repeats	
PolyQ family	17
Glycine histidine-rich	38
Collagen-like peptide	38
Other putative secreted peptides	224
Phenoloxidase inhibitor	1

*SVEP = *Simulium vittatum* erythema protein.

secreted polypeptides indicated in Supplemental Table S1 are abundant and complete enough to extract novel consensus sequences. Additionally, we have performed primer extension studies in several clones to obtain full- or near full-length sequences of products of interest. A total of 117 novel sequences, 72 of which code for putative secreted proteins, are grouped together in Supplemental Table S2.

Ubiquitous protein families of characterized function(s). *Enzymes.* Enzymes have been found in the salivary secretion of blood-sucking insects and ticks. In blood-sucking Nematocera, which also feed on sugar solutions, glycosidases such as maltase and amylases are commonly found. Enzymes can also function in antimicrobial functions, such as lysozyme and serine proteases that are possibly related to activation of the propheloloxidase cascade.[36,37] Endonucleases, together with hyaluronidases may also be present, possibly decreasing the skin viscosity and helping the diffusion of salivary pharmaceuticals and formation of the feeding cavity in pool-feeding insects such as sand flies and black flies.[13,38–43] Apyrase is also commonly found; it hydrolyzes host ATP and ADP to adenosine monophosphate (AMP), thus depleting these agonists of neutrophil and platelet aggregation.[6,7,44] The sialotranscriptome of *S. nigrimanum* reveals the presence of at least two glycosidases, possibly polymorphic as demonstrated by assembly of related coding sequences from multiple ESTs (Supplemental Table S1). Mosquitoes and *S. vittatum* sialotranscriptomes also reveal two glycosidases in their transcriptomes, consistently.

The salivary apyrase of blood-sucking arthropods is a typical case of convergent evolution. The salivary activity derives from at least three different genes. In mosquitoes and triatomine bugs of the genus *Triatoma*, members of the 5′ nucleotidase family have been co-opted for this function.[45–48] In sand flies and bed bugs a new gene product was characterized, in the form of the *Cimex*-type apyrase family.[49,50] In fleas it has been proposed that the activity derives from the CD39 protein family.[51] Previous sialotranscriptome analysis of *S. vittatum* found three ESTs coding for members of the 5′-nucleotidase family, and it was shown that the carboxy terminus of these putative enzymes lacked the inositol-phosphatide anchoring domain[46,52,53] typical of membrane-bound enzymes of this family, a situation previously shown in the salivary apyrase of the mosquito *Aedes aegypti*. The sialotranscriptome of *S. nigrimanum* reveals 19 ESTs associated with at least two different genes, allowing deduction of two 3′ truncated CDS, Sim-187, which is 76% identical to its *S. vittatum* homolog, and Sim-214, which is only 37% identical to the same protein in *S. vittatum*. Interestingly, while Sim-187 produces matches to mosquito and tabanid salivary apyrases, Sim-214 produces matches to *S. vittatum* and to bacterial proteins. Alignment of the simulid, mosquito, and bacterial proteins producing these best matches (Figure 2A) allows production of a bootstrapped phylogram showing strong support for 4 clades (Figure 2B); clade I contains the mosquito proteins; clade II has Sim-187 and the *S. vittatum* protein, joining the mosquito clade with 55% bootstrap support; clade III has four of the bacterial proteins; and strangely enough, clade IV contains Sim-214 together with the *Desulvovibrio piger* enzyme, with 81% bootstrap support, suggesting that Sim-187 may have arisen in *S. nigrimanum* by horizontal transfer (HT), as it has been proposed for some salivary gland expressed genes from mosquitoes.[10,12] However, the fact that Sim-214 produces a best match to the *S. vittatum* protein, which in turn produces best matches to mosquito salivary apyrases suggests more conservatively that Sim-214 arose by gene duplication from a common fly ancestral gene that evolved divergently from its parent to now resemble a bacterial protein.

The carboxy terminal regions of *Simulium* salivary proteins of the 5′-nucleotidase family were aligned with mosquito and tabanid (Chrysoptin) salivary apyrases, and with 5′-nucleotidases of human, mouse, cow, and *Drosophila* (Figure 3), showing that all salivary apyrases of blood-sucking Diptera lack the hydrophobic terminal region where the inositol anchor is located, indicating these salivary enzymes to be secreted and not membrane bound.

Comparison of the number of ESTs coding for members of the 5′-nucleotidase in the *S. nigrimanum* and *S. vittatum* sialotranscriptomes is informative. The previously described sialome of *S. vittatum*, a set of 1,483 ESTs, had only three sequences coding for this type of enzyme,[15] whereas in *S. nigrimanum* 13 ESTs were found out of 1,204 ESTs. The χ^2 derived from these numbers is 8.5917 with a $P = 0.00337$, indicating a highly significant difference, in accordance with enzyme measurements of salivary apyrase of New World black flies that indicated higher enzyme activity in vectors of human onchocerciasis.[54]

Destabilase is an endo-ε-(γ-Glu)-Lys isopeptidase, which cleaves isopeptide bonds formed by transglutaminase (Factor XIIIa) between Gln γ-carboxamide and the ε-amino group of lysine. This enzyme activity leads to dissolution of stabilized fibrin. Destabilase was first described in the salivary glands of the leech *Hirudo medicinalis*,[55] later shown to be the product of a multigene family that is related to the lysozyme superfamily.[56,57] Five ESTs in the *S. nigrimanum* sialotranscriptome have the destabilase CDD motif, possibly coding for two alleles. Whether these proteins function as the leech proteins do or as classical lysozymes remains to be determined.
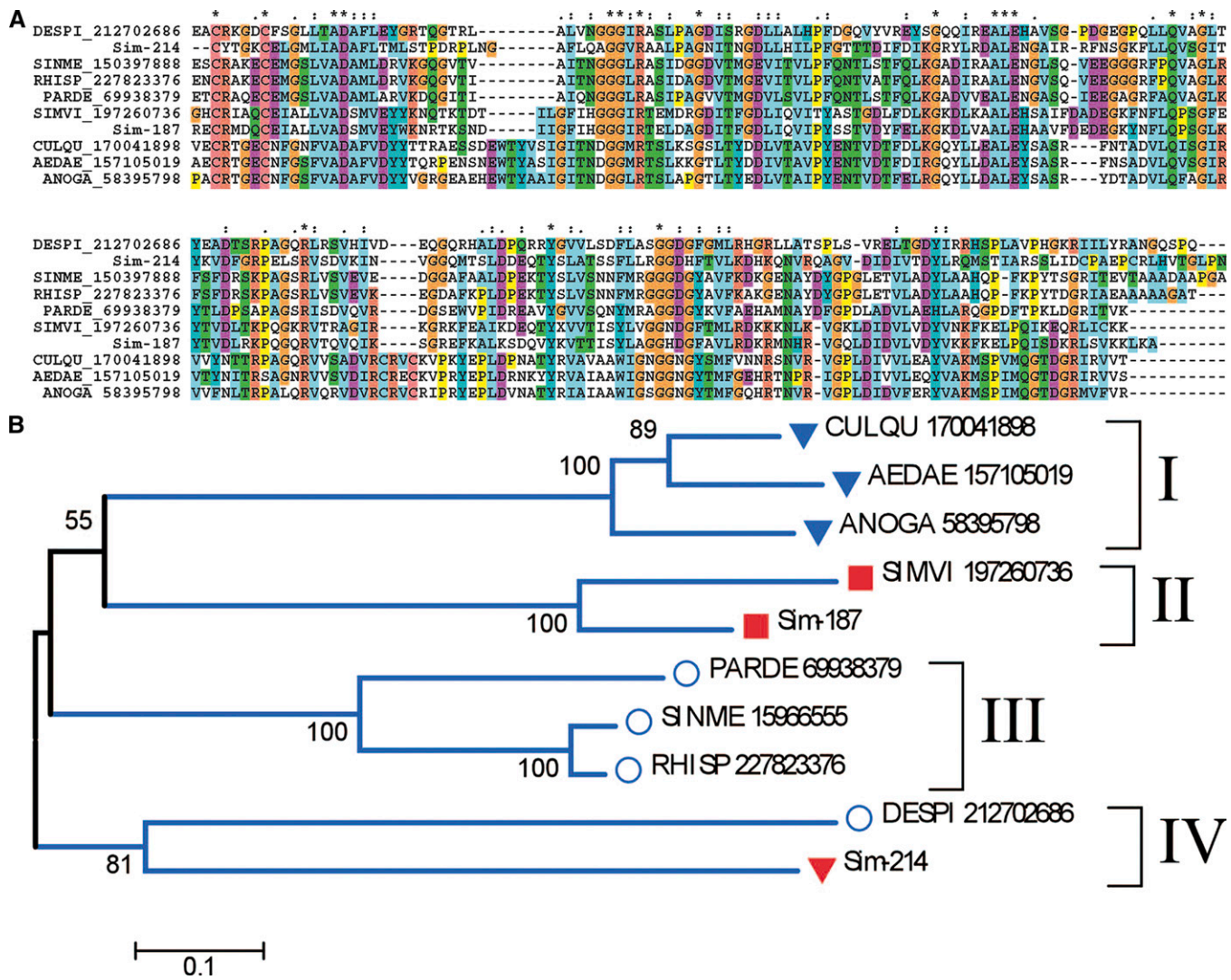
FIGURE 2. The 5′-nucleotidase/apyrase family of black flies, mosquitoes, and prokaryotes. (**A**) Clustal alignment. The *Simulium nigrimanum* proteins are Sim-214 and Sim187. The remaining sequences are named with the first three letters from the genus name followed by two letters from the species name and by their National Center for Biotechnology Information (NCBI) protein accession number. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. (**B**) Neighbor joining bootstrapped phylogram of the alignment in **A**, showing 4 clades. The numbers on the branches represent the percent bootstrap support. The bar in the bottom represents 10% amino acid divergence. Mosquito proteins are indicated by a triangle, *Simulium* proteins by a square, and bacterial proteins by a circle. For more details, see text. This figure appears in color at www.ajtmh.org.

Eight serine proteases sequences, four of which are full length, are presented in supplemental file S2, possibly deriving from 6 different genes. Sim-192 and Sim-193 appear to be alleles, whereas Sim-166 and Sim-167 appear to be differential transcripts of the same gene, or the product of recent gene duplications. These enzymes may function in immunity-related protease cascades or in blood feeding, such as in a fibrinolytic function.

Immunity-related gene products. Within this group, supplemental file S2 presents the full-length sequences of an antimicrobial peptide of the cecropin family, two lysozyme sequences (possibly alleles), and two members of the Gram-negative bacteria-binding protein, which function as bacterial recognition molecules that initiate immunity reactions. We also found a single EST matching a mosquito protein annotated as phenoloxidase inhibitor, and 54% identical to the *Musca* tyrosinase inhibitor deposited at SwissProt. Sn-307 has the

sequence VGD flanked by cysteines (CVRVGDWC), which is similar to snake disintegrins that block the adhesion of alpha4beta1 integrins.[58,59] Accordingly, Sn-307 may function as an inhibitor of vascular adhesion, as their snake venom counterparts.

Kunitz-domain containing peptides. The ubiquitous Kunitz domain is associated with proteins having serine protease inhibitor activity,[60,61] and ion-channel inhibitory activity.[62–65] Many tick anticlotting peptides from this family were described.[66–68] No transcripts coding for Kunitz domains have been found in transcriptomes of mosquitoes or sand flies, where anti-clotting activity was associated with serpins[69] or the anopheles-specific proteins of the anophelin family.[70,71] Kunitz peptides, however, were found in the sialotranscriptomes of the biting midge *Culicoides sonorensis*[72] and *S. vittatum*, but none of these proteins have been functionally characterized. Figure 4 shows the full-length sequences of four Kunitz
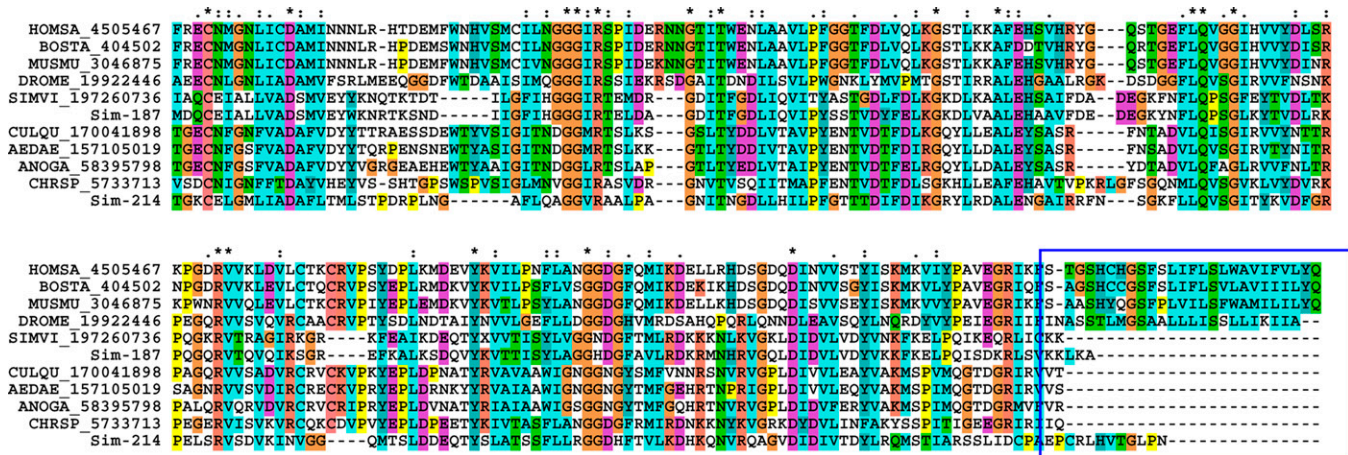
FIGURE 3. The 5'-nucleotidase/apyrase family of black flies, mosquitoes, tabanids compared with *Drosophila* and mammalian enzymes. The *Simulium nigrimanum* proteins are Sim-214 and Sim187. The remaining sequences are named with the first three letters from the genus name followed by two letters from the species name and by their National Center for Biotechnology Information (NCBI) protein accession number. The box shows the deletion of the membrane anchor region in the salivary apyrases of Diptera. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. This figure appears in color at www.ajtmh.org.

domain-containing peptides from *S. nigrimanum*, deducted from the assembly of 7–41 ESTs, indicating these transcripts are relatively abundant. Alignment of the four *S. nigrimanum* proteins with the two Kunitz proteins described from *S. vittatum* clearly shows two polypeptide families (Figure 4A), one of them containing an extended carboxy terminal domain populated by basic peptides, reminiscent of the basic tail family of tick salivary proteins. This basic tail may lead the peptides

to associate to negatively charged phospholipids such as those on the surface of activated platelets that serve as a matrix for assembly of blood coagulation proteolytic complexes such as the Xase and prothrombinase complexes.[73–75] The phylogram of the alignment (Figure 4B) shows strong support for 2 clades of *Simulium* proteins (marked I and II), indicating that the two black flies shared a common ancestor already containing the two different genes. The phylogram also indicates that
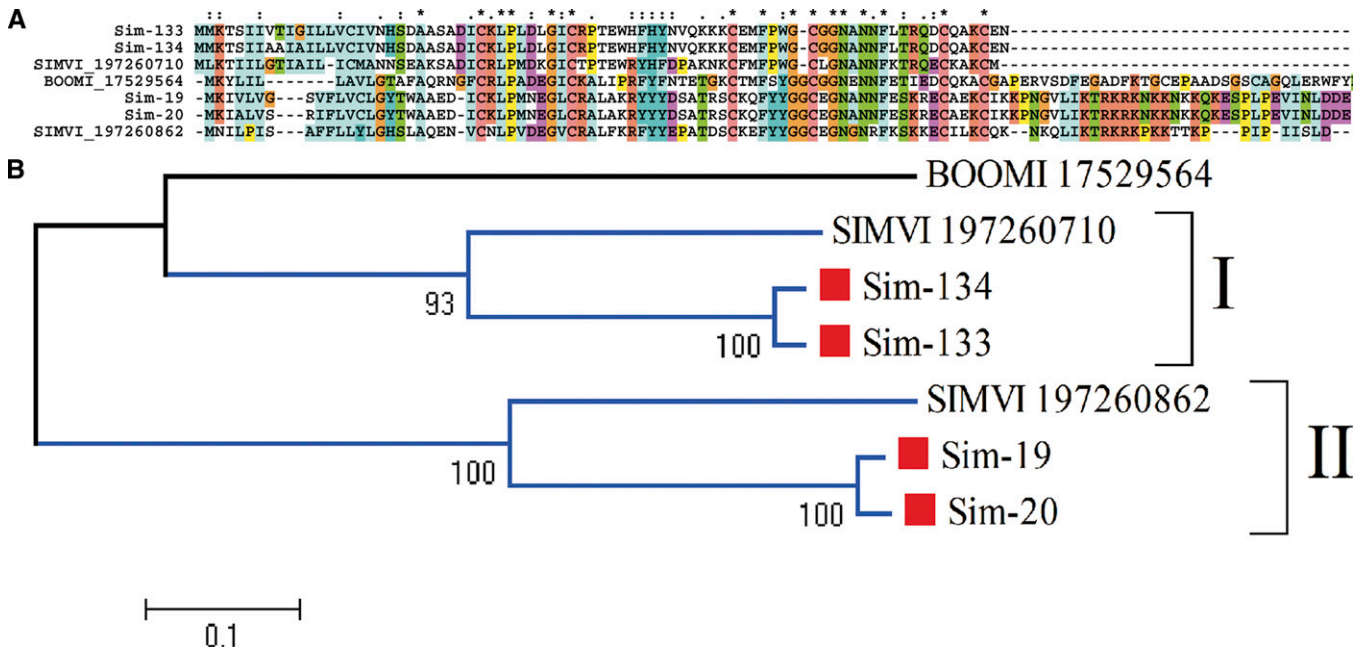
FIGURE 4. The *Simulium* Kunitz salivary protein family. (**A**) Clustal alignment of 4 *S. nigrimanum* proteins (starting with Sim-) with 2 *Simulium vittatum* proteins, indicated by SIMVI_X where X is their National Center for Biotechnology Information (NCBI) accession numbers. The tick anti-thrombin peptide named boophilin (BOOMI_17529564) is included as an out-group. The symbols above the alignment indicate identity (*), high similarity (:), and similarity (.) of residues in the indicated alignment position. (**B**) Phylogram derived from the alignment in **A**. *Simulium nigrimanum* sequences are marked with a square. The numbers on the tree bifurcations indicate the percentage bootstrap support above 50%. The bar at the bottom represents 10% amino acid substitution. Protein sequences were aligned by the Clustal program[39] and the dendrogram was done with the Mega package[41] after 10,000 bootstraps with the neighbor joining (NJ) algorithm. For more details, see text. This figure appears in color at www.ajtmh.org.

S. nigrimanum has two genes in each clade, or more probably each has at least two common alleles.

Antigen 5 protein family. AG5-related salivary products are ubiquitous in sialotranscriptomes of blood-feeding arthropods, being members of a group of secreted proteins that belong to the CAP family (Cys-rich secretory proteins; AG5 proteins of insects; pathogenesis-related protein 1 of plants).[76] The majority of these animal proteins have no known function. The notable exceptions include proteolytic activity in Conus,[77] smooth muscle-relaxing activity in snake venoms,[78,79] and salivary neurotoxin activity in the venomous lizard *Heloderma horridum*.[80] Recently, an antigen-5 protein from the saliva of a tabanid fly[81] was shown to inhibit platelet aggregation by the unusual acquisition of a typical RGD domain that is known to prevent fibrinogen binding to platelets and its ensuing aggregation,[82] and a stable fly salivary protein of this family was shown to bind host immunoglobulins and may function as an inhibitor of the classical pathway of complement activation.[83] These diversity of functions prevent generalizations regarding this protein family. The sialotranscriptome of *S. nigrimanum* produced 24 ESTs matching members of this family, which produce best matches to salivary proteins of *Culicoides,* sand flies, and mosquitoes.

Protein families exclusive to blood-feeding Diptera. *D7/ odorant-binding protein (OBP) family.* The D7 salivary family is typical of blood-feeding Nematocera.[84] It is related to the OBP family,[85] but containing two additional helices per domain.[86] Members of this family have been found in all mosquito, *Culicoides*, sand fly, and black fly transcriptomes so far studied. Single- or double-domain proteins exist, constituting the short and long D7 subfamilies.[87] *Anopheles gambiae* has. Eight genes encoding these proteins, with 3 genes encoding the long forms and 5 encoding the short forms.[12] Mosquito proteins have recently been shown to bind biogenic amines[88] and inflammatory lipids[89] thus helping blood feeding by sequestering vasoconstrictory, inflammatory and platelet aggregation agonists. In *S. vittatum* both short and long D7 proteins were also found,[15] one of which was demonstrated in a patent to be an anti-clotting agent.[90] The *S. nigrimanum* sialotranscriptome reveals both long and short forms of the D7 family, with 34 deduced protein sequences being available in supplemental file S2, many of which appear to be alleles. This family is also the most represented in the sialotranscriptome, with a total of 317 EST of the total of 1,255 attributed to secreted products (Table 3).

Phylogenetic analysis of the D7 sequences of *S. nigrimanum* combined with those from *S. vittatum* reveals 9 clades with strong bootstrap support (supplemental file S2 and Figure 5). Some of these clades, such as I, IV, V, VII, and IX have closely related sequences that could be alleles, whereas others such as II, III, VII, and VIII contain sequences with over 15% amino acid divergence indicative of two or more genes. Accordingly, at least 13 genes must exist in *S. nigrimanum* coding for D7 proteins, possibly 14 if we count 3 genes in Clade II. Notice also the multispecies clades III, VI, and VIII, indicating the ancestor of *S. vittatum* and *S. nigrimanum* already had these genes. The phylogram also indicates that the majority of the short sequences coalesce with strong bootstrap support under one super clade, named Short in Figure 5, containing clades I–V. Long D7 sequences cluster both with other long sequences, but also with short *S. nigrimanum* sequences. It is also to be noticed that the two related short sequences in clade VII,

Sim-188 and Sim-189 have a cluster of basic AAs in their carboxy terminal region that might lead these proteins to associate with negatively charged lipids that are important for assembly of blood clotting enzymes.[73,91] Finally, the *S. vittatum* sequence gi|197260866, the salivary anti-thrombin of *S. vittatum*,[90] localizes in Clade III, suggesting the homologous proteins of *S. nigrimanum*, which are 67% identical in AAs sequence, may also display anti-thrombin activity.

30 kDa antigen/Aegyptin family. A salivary immunogenic protein from *Ae. aegypti* named 30 kDa antigen[92] is similar to anopheline proteins named GE-rich protein for its abundance in glycine and glutamate residues. Recent functional analysis of members of this protein family from *Aedes* and *Anopheles* revealed they are antagonists of collagen-induced platelet aggregation.[93,94] Alignment of black fly and mosquito proteins (Figure 6) shows the previously described distinct domains,[93] including the signal peptide region, a glycine/aspartate/glutamate-rich region, and a relatively more conserved carboxyterminus, where the conserved pattern T-Y-x(6)-L-x(19,22)-Q-x(18,19)-I-x(2)-C-F-x(20)-C-x(3,10)-C-x(20,21)-C is found. This unique family supports a common origin of hematophagy between black flies and mosquitoes, as proposed by Grimaldi and Engels.[95]

Hyp16 family. The first member of this family was identified in the sialotranscriptome of *Anopheles. stephensi*,[14] followed by identification in the sialome of *Aedes albopictus*,[11] and more recently, in the sialome of the stable fly, *Stomoxys calcitrans* where it was found abundantly expressed.[41] Sim-692 is 29% identical and 48% similar in its sequence to the *An. stephensi* protein. Homologues in *An. gambiae* and *Ae. aegypti* are found in the deducted proteomes of these mosquitoes. Alignment of these protein sequences reveal the conserved pattern G-x(12,13)-C-D-x(3)-C-P-x(5)-C-x(3)-K-x(12,15)-C-x(4)-G-x(4)-K-x(17,19)-E that might help to identify members of this family, none of which have known function (Figure 7).

Protein families exclusive to black flies. *SVEP.* The salivary vasodilator of *S. vittatum*, named SVEP for *S. vittatum* erythema protein, has been previously identified as a novel protein[96] and the recombinant protein expressed and functionally characterized as a vasodilator possibly activating ATP-dependent K+ channels.[97] The sialotranscriptome of *S. vittatum* identified SVEP to belong to a diverse multigene family containing at least 5 genes.[15] As expected, *S. nigrimanum* has homologues to SVEP, varying from 45–66% identity. Alignment of the *Simulium* SVEPs shows similar sized sequences with relatively few conserved AAs spread over the length of the sequence (Figure 8A). The phylogram also indicates that each species has at least 5 genes coding for members of the family, and that their products group within their own species, even though they can be very distant from each other, showing a *S. vittatum* and a *S. nigrimanum* clade, and an intermediate *S. nigrimanum* clade that might derive from a gene closer to the ancestral gene (Figure 8B). This scenario indicates either that the gene duplications in each species occurred after the split from a common ancestor, or that gene conversion events within each species occurred to keep some degree of homogeneity among the genes.

To attempt an insight into the origins of the SVEP family, members of the *S. nigrimanum* proteins were submitted to PSI-BLAST querying the NR protein database from NCBI. As expected, the protein Sim-116, which belongs to the canonical SVEP family, retrieved members of the *S. vittatum*
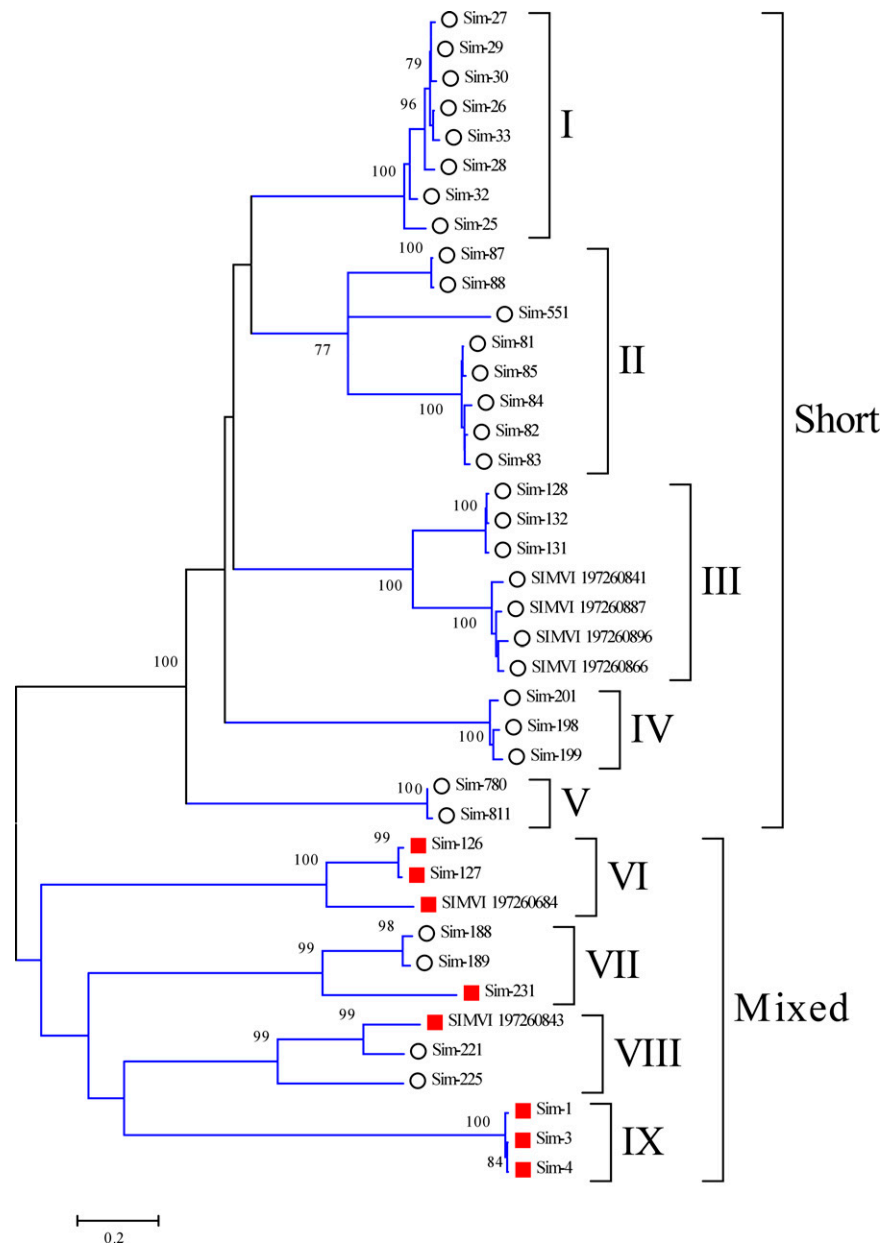
FIGURE 5.   The *Simulium* D7 salivary protein family. Phylogram deriving from the alignment of *S. nigrimanum* proteins (starting with Sim-) with those of *Simulium vittatum*, indicated by SIMVI_X where X is their National Center for Biotechnology Information (NCBI) accession numbers. The numbers on the tree bifurcations indicate the percentage bootstrap support above 75%. The bar at the bottom represents 20% amino acid substitution. Members of the short D7 family are marked with a circle, those from the long family with a square. Protein sequences were aligned by the Clustal program39 and the dendrogram was done with the Mega package41 after 10,000 bootstraps with the neighbor joining (NJ) algorithm. For more details, see text. This figure appears in color at www.ajtmh.org.

SVEP family, but in the second round it retrieved a bacterial insect toxin with low e-value. Further iterations led to retrieving many additional bacterial proteins annotated as glycosidases and ricin, suggesting the SVEP family originated from a carbohydrate domain-containing ancestor. Although C-type lectins are a common finding in sialotranscriptomes of mosquitoes and sand flies, a ricin member in a sialotranscriptome is a novelty.

Collagen-like family. This protein family was previously identified in the *S. vittatum* sialome. It consists of basic proteins (pI = 10) rich in glycine, lysine, and proline, with over 20% of the protein consisting of Pro+Gly. No cysteine residues are found in the mature proteins. The amino terminal region has a relatively conserved region (Figure 9A), but the carboxy terminus region determines at least three subfamilies, as follows: 1) there are two long *S. vittatum* proteins, responsible for the insert/gap regions seen in Figure 8A, and marked as clade Long in Figure 9B. 2) Two shorter *S. vittatum* proteins possess a basic tail of lysine-rich residues that might drive these proteins to binding anionic phospholipids[73]; these are marked as Basic tail clade in Figure 9B. The *S. nigrimanum* sequences form a single uniform clade (Figure 9B), and their

```
AEDAE_94469546   -----MKPLVKLFLLFCLVG-IVLSRPMPEDEEPVAEG-GDDDASG-ESEGEEETTDD------AGGDGGEEENEGEEHAGDKDAGGEDTGKEENTGHDDAGEEDAGEEDAGEEDAGEEDAGEEDAEEKEEGEKEDAGDDAGSDDGEEDST
AEDAL_56417504   -----MKPLAKLFMVFCLIG-IVLSRPMPEDEEGEEESPDDASGDETEGGEEKTDDG------ASEDGGEESKADEEDGGENADGEDAGGDDAGGENADGENTDGENTDGENTDGE-NADGEDAE---GSK-DEGDDSEGDGSKEEST
AEDAE_18568322   -----MKYLLTFLMALSLVN-LMLTRPTPEDD-------------------------------GGTSEEPQTQETTGSDEKNG------------------------------ASEEPNADDASKPDDVEEKGDDDTAKKEDDGESK
AEDAL_56417500   -----MRSLLVLFAALCLVG-FVLAKPIPEDDT----------EATTEEPKAEDAAEQT---------------------------TAEDAGGSDGEKND--DEKPSDDAANDQPKDVE
AEDAL_56417508   -----MQLFPISLVVSCLVFSFVLSLPLPEE--------------ASTSEETEATTEAESA---------------------------AADETTAT---------GADDAESGSSEKNNE
ANOST_37201975   -----MKLLLLLASVLCLAL-IVSARPSDETTDQEESSTELSEDTSA--------SYHQEEDTEETGAEDTGDGSSELESSSEGHEDGSEDATGEEGGAGKKGEAGEEDAEGA-EGEGEEGAGE
ANOFU_114864639  -----MKSLLLLASVLCLAL-IVSARPSDETTPDQETSTELSEDTSA---------DDTQEEETVDTGAES---DADSES-GADAEG---GDASETEDTGTGESEASPD--EEHTDEKASDDAGEGTDDTKDEDTEGHDD
ANOGA_30_KDA     -----MKFLLLVASVLCLVL-IVSARPADDTSDQESSTELSEDAG----------------AEEGAEDAGSDA---EADAGAAD--GEEGATDTEGAEGDDSEMDSAMKEEGEEGAGSDDAVS--GADDETEESKDDAEED
ANODI_31982968   -----MKFLLLLAGVLCLAL-IVTARPQDEEAI-ETTTQLSEDAG---------EGTHEEGDBEEHSDS---EAGGSKGDEEGEEGGEEDEVSDSHDGADEEEEHSEGDD-AGGDDATS--EDAEEEGEGDAGESDD
ANOAL_71389021   -----MKLYLLVSVLCLAL-IVAARPNEDTTE-ETSALGPEDGDA-----------------SDAPRDTPEDTTDE---------------------------GDAPTSEEHDMLTSDDEVEDHGKESE--NGDDTADQTDEAESTD
Sim-50           MKKFPCLIPLLLALTALHVKGSPIPDEESEETEDAEQEEDAEEREHAEEEAQEDGDQTQDTTQESDDGQEETANQGSDEQQREDGSHGTPETDDEAAGKRTSDVEGGDQNDGEENGSQENGEEVTGGEDDHGAQDSEGGEEDAAAGEA
Sim-54           MKKFPCLIPLLLALTALHVKGSPIPDEESEETEDAEEEEDAEEREHAEEEAQEDGDQTQDTTQESDDGQEETVNQGSDEQQREDGSHGTPETDDEAAGKRTSDVEGGDQNDGEENGSQENGEEVTGGEDDHGAQDSEGGEEDAAAGEA
Sim-55           MKKFPCLIPLLLALTALHVKGSPIPDEESEETEDAEHEEDAEEREDAEEEAQEDGDQTQDTTQESDDGQEETANQGSDEQQREDGSHGTPFTDDEAAGKRTSDVEGGDQNDGEENGSQENGEEVTGGEDDHGAQDSEGGEEDAAAGEA
Sim-49           MKKFPCLIPLLLALTALHVKGSPIPDEESEETEDAEHEEDAEEREDAEEEAQEDGDQTQDTTQESDDGQEETVNQGSDEQQREDGSHGTPFTDDEAAGKSNADRVEGGDQNDGEENGSQENGEEVTGGEDDHGAQDSEGGEVDAGAGEA
SIMVI_197260700  ---------EETDDQGS--VDQEEEETDDQDSRTDDDSVESSRTQGSGVTHDDVASDRDGTESDLEDDQESHTQDDEENDRQDSEEQDDTDG---RGGHAGAADEEDSVGGLGQDNRDGQVTDSDAEGREVEGAGARGGKRGKGA
CULPI_38350685   -----MKLVRPTLLALFLLIGILANAEAETETLPSTESFPS---------------------SAATEDDQPEKPTNNS--------------------EPSSHPAEDDDDQPKEPTKDGESSEPDDNDDDSEASSEKP
```

```
                                :  :  **.:.*.            *.                                  .:.* . * .**.:.           *            .
AEDAE_94469546   --------GGDEGEDNAEDSKGSEKNDPADTYRQVVALLDKDTKVDHIQSEYLRSALNNDLQSEVRVPVVEAIGRIGDYSKIQGCFKSMGKDVKKVISEEEKFKSCMSKKKS-EYQCSEDSFAAAKSKLSPITSKIKSCVSSKGR---
AEDAL_56417504   --------GGDEGGDNAGGEGGGGEHDPVNTYHKVVALIDKDTKVDNIQSEYLRSALNNDLQSEVRNPVVEAISRLGSFSKIEGCFKSMGSDVKKVIDEEQKAFKDCMTKKKS-EYECSEDSFASAKGKLSPITSKIKSCVSSKGQ---
AEDAE_18568322   --------DGEGSEKSDKE-KGEPKNDPRETYNKVIEQLDQ-IKVDNVEDGHERSELAADIQRYLRNPIVDVIGSAGDFSKIAKCFKSMVGDAKKAIEEDVKGFKECTAKKDSNAYQCSQD-RSTVQDKIAKMSSKIASCVASNRS---
AEDAL_56417500   --------GGEDGKESPKT-EGEAKNDPQETYNKVIELLDQ-IKVDNVENGYIERSELTSDLQTYLRNPIVDAIGTVGDFSKIADCFKSMGDAKKAIEEIELKAFKECTDKKESDAYQCSKN-HSTVQGKIAKISSTINSCVVSKRA---
AEDAL_56417508   --------GGETSAPEAKE-EGSGKDDRLDTYNKVVELLDKDTGVNNVENGFLRAFLDNNLQSHKRVPVIDAIGLVGDFSKIAGCFTSMGADVKKIVDDKLFSPTTCKSGKEAGDAKCSED-GSSVGGQFDQITSKIKECVSSKGKGK-
ANOST_37201975   EGGAGEEGGADEEGSAGEEGGAGEEGGEESPRNTYHQVEHLLKNIMVG-TKNNNLKFSFILARLQERLMNPTIDLVGSISYISKIKECFDSLADDVKSLVEKSETSYEECSKDKN--PHCGSEGTRELDEGLIEREQKLSDCIVEKRDSE-
ANOFU_114864639  ---ASEEGADGEGAGEDQQGGGGEGGEESPRNTYRQVHKLLKNIMKVG-TKDKYLKSFVLARLQERLMNPTIDLISSVQKYISKIKECFSSLAKDVGSLVKESDMSYEECSKDKSN--PNCGSEGTRELDEGLIEREQKLSDCIVEKRDSE-
ANOGA_30_KDA     --BEEGGEEGGDGASGGEGGE--KESPRNTYRQVHKLLKKIMKVD-TKDYLKLKSFVVGRLQERLMNPTIDLVSTEKYISKSLDKDVSAMVKESESYEECSKDKTN--TSCGTEGTRELDDGLIEREQELSDCIVDKRDAE-
ANODI_31982968   ----SEEGGKESDAGAGGKGGGE--KDDRRNTYRQVHDQLKKIMKVG-TKDGYLKSFVVARLQERLMNPTIDLIGTIGKYSKIKECFSSLAKDVAALVKGSES SYEECTKDKTN--PSCGSEGTHDLDEGLVDRQQTLSDCIVEKRDAQ-
ANOAL_71389021   --GESDGSAEDGAGGSEGGKGGGKAGESPRNTYRQVHKLLKVMKVD-VKDKYLKSFIVARLQERLMNPTIDLVGINIEKYSKIKECFSSVEKDTKDLVKESEAAYKECSEDKSN--TSCGSEGTRELDDGLIDLEQELSDCIVEKRDADS
Sim-50           GGKKSKGKGAKGGQGSNRAGSSSSGGEESAAHTYQEVMRILNG-IKIN--AKEPLKSRLVNGLQY-LKVDTINRVENIQKFSRIEHCFGNLDHIVQLYAGDAVERNDYCKG-------TCIDMVSNEFRGQMEELTSRLYPCLQGGYY---
Sim-54           GGKKSKGKGAKGGQGSNRAGSSSSGGEESAAHTYQEVMRILNG-IKIN--AKEPLKSRLVNGLQY-LKVDTINRVENIQKFSRIEHCFGNLDHIVQLYAGDAVERNDYCKG-------TCIDMVSNEFRGQMEELTSRLYPCLQGGYY---
Sim-55           GGKKSKGKGAKGGQGSNRAGSSSSGGEESAAHTYQEVMRILNG-IKIN--AKEPLKSRLVNGLQY-LKVDTINRVENIQKFSRIEHCFGNLDHIVQLYAGDAVERNDYCKG-------TCIDMVSNEFRGQMEELTSRLYPCLKGATIK--
Sim-49           GGKKSKGKGAKGGQGSNRAGSSSSGGEESAAHTYQEVMRILNG-IKIN--AKEPLKSRLVNGLQY-LKVDTINRVENIQKFSRIEHCFGNLDHIVQLYAGDAVERNDYCKG-------TCIDMVSNEFRGQMEELTSRLYPCLQGGYY---
SIMVI_197260700  RGKKGRGK-STDAEGGGSAGSSYSGEESAAHTYQQVMQILNG-IQIN--ANEPLKSINGLQD-LKTQTINMVVHIEKFSRIEHCFGNLDYIVQLYAGDAVERNDYCKG-------ACIDVVSNDFRTQMNELSGRLSPCLQGGYY---
CULPI_38350685   ---------SKKGGKGPGVGKGPSQKGSYAATYQEIVKILNS-VDFKSIEDLHLQRMLENAFQARVRNPVLEQTGQIADFGAIKSCFGKLTGEVKKLINAAEKQCKTCKTGGGYS-SGCTDQQENAFADGVIDLATTLQGCISSKRKN--
```
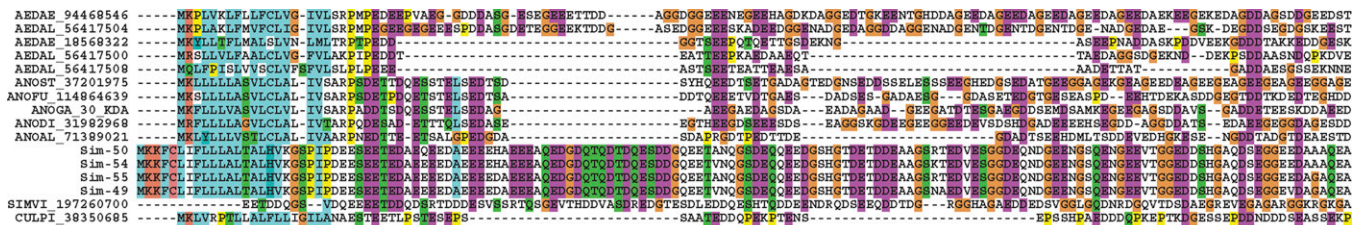
FIGURE 6. The 30 kDa antigen/Aegyptin family of black flies and mosquitoes. The *Simulium nigrimanum* proteins are indicated by Sim-XX where XX is the number indicated in supplemental file S2. The remaining sequences are named with the first three letters from the genus name followed by two letters from the species name and by their National Center for Biotechnology Information (NCBI) protein accession number. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. For more details, see text. This figure appears in color at www.ajtmh.org.

sequences are more homogeneous (Figure 9A). The *S. vittatum* sequence gi|197260678 does not belong to any of the two strong clades, representing a possible link between the two groups of sequences. Functional analysis of recombinant proteins from this family should take into consideration that they may be modified into hydroxyl proline and lysine, as collagen is modified.[98]

Acidic H P Q E-rich proteins of low complexity. This protein family was also identified in the *S. vittatum* sialotranscriptome, where it comprised its most abundant cluster of ESTs. These have low complexity and abound in histidine, proline and glutamic acid residues, with regions of Gly-His or Pro-His repeats. They are possibly related to mosquito and Culicoides proteins that also show Pro-His and Gly-His repeats. The *S. nigrimanum* proteins are ~65% identical in primary sequence to their *S. vittatum* homologues. It is possible that the His repeats may function as antimicrobials by chelating Zn or other trace element ions.[99–101]

Simulium mucins. Mucins are low complexity proteins rich in serine and threonine residues that can accept N-acetyl-galactosamine residues.[26] The *S. nigrimanum* sialotranscriptome reveals the presence of such proteins, over one-third of their residues consisting of Ser+Thr, and displaying 33–136 putative galactosylation sites.

Simulium basic 7–13 kDa family. The *S. nigrimanum* sialotranscriptome revealed seven protein sequences with two or more ESTs each coding for basic (pI 8.1–10.6) proteins ranging from 7 to 13 kDa in predicted mature MW. This group of proteins clusters when they are compared at 30% similarity over 50% of the smaller length, indicating they are possibly related. These seven sequences belong to two more closely related groups, one with four sequences, the other with three sequences. The second group contains sequences that are ~65% identical in primary sequence to a previously described orphan *S. vittatum* protein.

Sv 7.8 kDa family. A second group of six proteins also coding for basic proteins of size varying from 7 to 11 kDa can also be identified by their 30% similarity level. Members of this family match previously described *S. vittatum* proteins varying in identity from 40% to 57%, which were named Sv

```
                      .        :              :                  :  *   .  .:  :            **  :  **      : :*::  *  .  ..              :. *   .   *
ANOGA_58390829   MLSKGHVMMVALCAVCAIFGTTAVEGAALNRSPRQLSSLLTLSGESNARIEN-GTIICDTLKCPAESFKCVIVKNSTKDDVNKVQVTRECLDPAGKATA
ANOST_27372897   MLSKGQTLLLVVFAV-AVFGTVSVEGAALNRTPRQLSSLLTLSGESNSRIEN-GTILCDTLKCPPESFKCVIVKNNSTKDDVNKIRVTRECLDPAGKATA
AEDAL_56417416   -MMNSKQVILSLGLLLVVVGLSGVQGAAL-REKRQLNSLLTLSGEANARIEN-GTIFCDTLKCPDASFKCIIVKNNTKEDVNVEVTRECLDAAGKATA
CULQU_167878877  --MNSQQMILCFG-LVLAVGLSLADGAVV-REKRQLNTLLTISGESNARIVN-GTIFCDTLKCPDASFKCIIVKNNTKEDVNVEVTRECLDAAGKATA
STOCA_63148846   MAIRTSLVFLLASFVCSAYSQ-----------------LLTPIGRANIEITD-RRFQCDRIECPMETERCVVTKEKHPQSDWMLVVSNQCYSQDGTLLK
STOCA_63148848   ----MKIFAILLAVSYISSAYG----------------FSNIGRANVEISD-SKFQCDKMLCPSSAERCVVTKEKDPKNPVRVLVRTNTCLSRTGVELR
Sim-692          MGKQFVTLWAIATLIGLAVGQANYG----------NLIASNIGAMTLQVYDTGIIINCDNFICPKSTYKCIVDKITDEKS---TRITKSCHDANGQIVD
```

```
                 *             .  .  :            :  *                                          .                          ::  ..                      .
ANOGA_58390829   KTVTTEEASNFPGQQFE-SYAEIDKNGNIASFDN--RGNTYNHSGTGDIHS---------YLYEQIEDLHKAILDQLSNAGNPFGA---------------
ANOST_27372897   KNVTTEETSSFPGQQFE-SYAEIDKNGNIASFDN--RGNTYNHSGSGDIHS---------YLYEQIEDLHKAILDQLTNAGNPFTAQ--------------
AEDAL_56417416   KTIEREEANQFPGTHFE-SYAEIDKNGNIASFDN--RGNSYNHSGSGDLHS---------YLYKQIDELHKQILDQLQGVSSSYRIP-------------
CULQU_167878877  KTVEKEEPNQFPGGKYE-SYAEIDKDGNIASFDN--RGNSYNHSGSGDVHS---------YLYQQIEDLHKSILDQLTNAGTPYRIP-------------
STOCA_63148846   KAEKQQ-SIDPKVHVK-IHLEGRRIYG---------DRPSVGENDAD---------------DDEGDDAFNRAVDELSKDFDLN-----------------
STOCA_63148848   KKTWQE-GTNPNAKVS-VHIEGLRYVGYTPVKP--AANGKPAKPGNPAKTKPQKPNSNWNLPDSNDDDNDAFNRAVDELAKDFDF----------------
Sim-692          KEEADIPNLTSETNVRRSYAEADIHGRVVVKNHGYGGDNFNYQQTGSVGN-QLTPEDIERLRRLQQDINDYTIRLQQSIQSQVQDTLNRIRNMQFPFVR
```
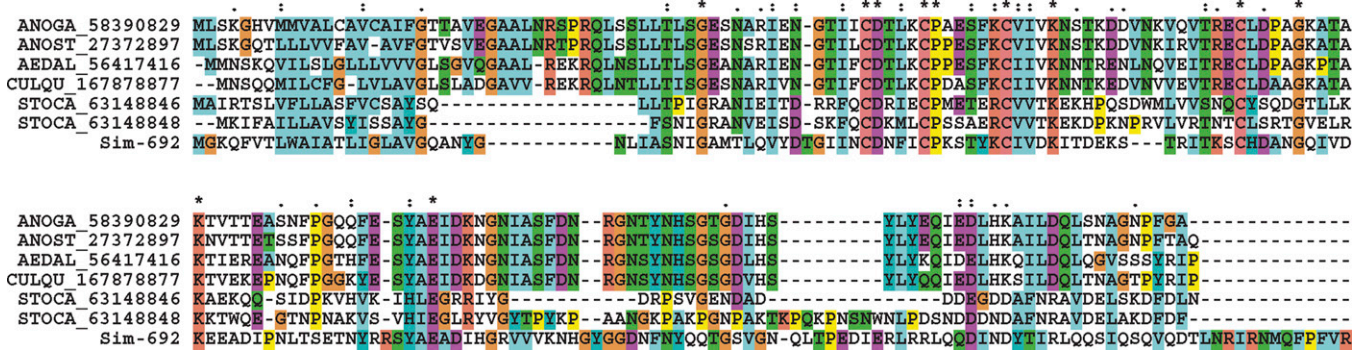
FIGURE 7. The Hyp16 family of blood-sucking Diptera. The *Simulium nigrimanum* protein is indicated by Sim-692. The remaining sequences are named with the first three letters from the genus name followed by two letters from the species name and by their National Center for Biotechnology Information (NCBI) protein accession number. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. For more details, see text. This figure appears in color at www.ajtmh.org.
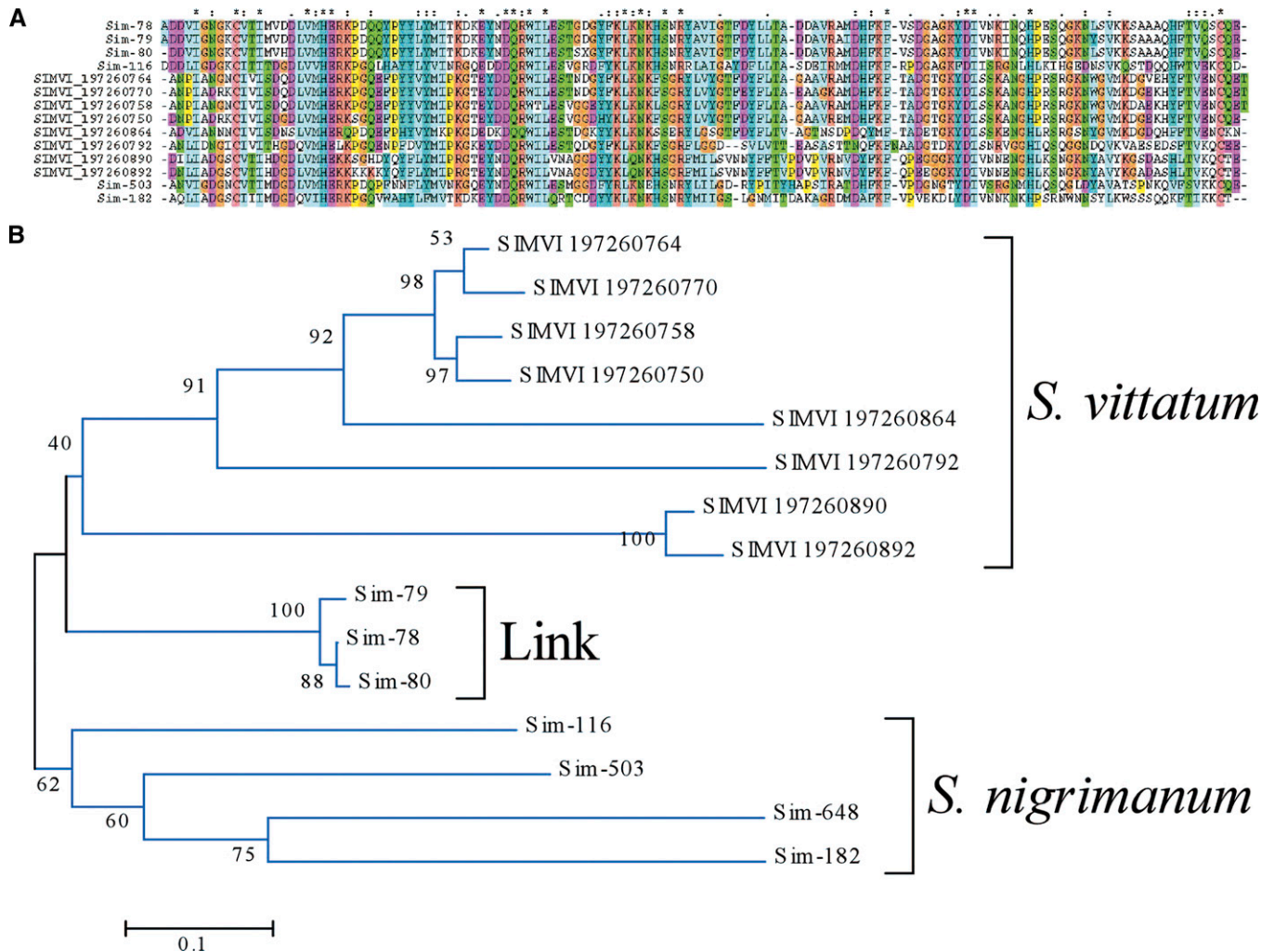
FIGURE 8. The *Simulium vittatum* erythema protein (SVEP) superfamily of *Simulium*. (**A**) Clustal aligment. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. (**B**) Bootstrapped phylogram. The *Simulium nigrimanum* proteins are indicated by a square. The *S. vittatum* sequences are named SIMVI followed by their National Center for Biotechnology Information (NCBI) protein accession number. The vasodilator SVEP is marked with a triangle. The numbers on the tree bifurcations indicate the percentage bootstrap support. The bar at the bottom represents 10% amino acid substitution. Protein sequences were aligned by the Clustal program39 and the dendrogram was done with the Mega package41 after 10,000 bootstraps with the neighbor joining (NJ) algorithm. For more details, see text. This figure appears in color at www.ajtmh.org.

7.8 kDa family previously. Other members of this group of *S. nigrimanum* proteins match *S. vittatum* proteins classified as orphans, indicating this group may derive from a very divergent multifamily gene from *Simulium*. The 7.8 kDa family clusters at the borderline 25% similarity level with the above described 7–13 kDa family, but no conserved AAs can be identified in the alignment, indicating if these two families share a common ancestor, the family has diverged beyond recognition.

Basic 13 kDa family. Five deducted protein sequences from the *S. nigrimanum* sialotranscriptome match a previously described orphan protein from *S. vittatum*, constituting a new protein family here named the basic 13 kDa family. They do not match significantly any other known protein.

Sv 7 kDa family. Five sequences from the *S. nigrimanum* sialotranscriptome match two smaller sequences from the *S. vittatum* sialome previously named the Sv 7 kDa family. The *S. nigrimanum* sequences have a glycine-rich insert when com-

pared with the *S. vittatum* relatives (Figure 10). This family is Ser+Thr-rich, and have eight potential N-acetyl-galactosylation sites.

Sv 4.8 kDa family. This *Simulium* family produces shorter peptides in *S. nigrimanum* then in *S. vittatum* (Figure 11). All mature peptides are devoid of cysteines.

Other deorphanized Simulium *proteins.* Three additional protein families were found in common between *S. vittatum* and *S. nigrimanum*, and no other known protein. These constitute the 5 Cys, the basic 13 kDa and the basic 7 kDa family. These proteins do not produce significant similarities to other proteins in the NR database.

Families exclusive to *S. nigrimanum*. Five distinct protein families, all with two or more ESTs, were found in the *S. nigrimanum* sialotranscriptome, plus seven additional orphan proteins, as follows:

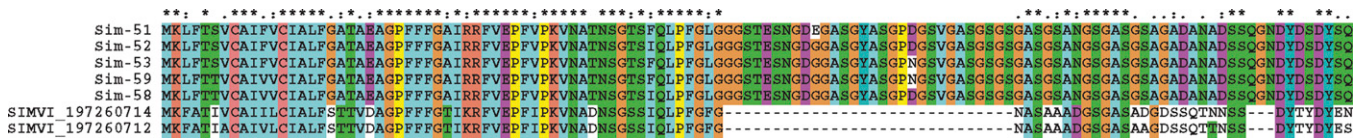*Sn actinohivin-like.* Sim-177 and Sim-178 have 32% and 31% identities to an actinomycete protein named Actinohivin,

A

B

0.05

FIGURE 9. The collagen-like superfamily of *Simulium*. (**A**) Clustal aligment. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. (**B**) Bootstrapped phylogram. The *Simulium nigrimanum* proteins are indicated by Sim- and their identification number. The *Simulium vittatum* sequences are named SIMVI followed by their National Center for Biotechnology Information (NCBI) protein accession number. The numbers on the tree bifurcations indicate the percentage bootstrap support above 75%. The bar at the bottom represents 5% amino acid substitution. Protein sequences were aligned by the Clustal program39 and the dendrogram was done with the Mega package41 after 10,000 bootstraps with the neighbor joining (NJ) algorithm. For more details, see text. This figure appears in color at www.ajtmh.org.

and lesser identities to other bacterial proteins annotated as xylanases. Actinohivin is a lectin with anti-human immunodeficiency virus (HIV) properties.[102] This suggests the fly proteins to be possible sugar ligands, which is further suggested by their CDD match to the ricin motif. Sim-177 and actinohivin blast alignments were used to build a search model for Psiblast, which produced matches to proteins annotated as glycosydases or lectins, including ricin, with small e-values on the second round, further suggesting these black fly proteins belong to the ricin lectin family of proteins. These results suggest that the

Sn actinohivin-like family may be related to the SVEP family, having evolved beyond recognition at the primary structure level.

The similarity of this family to actinomycete proteins suggests it derives from a microbial contaminant, or horizontal transfer, or convergent evolution. Because a total of nine ESTs were obtained for this family, it would be expected that many more microbial contaminants would be found in the assembled database. Confirmation of these sequences at the fly's genome level may help to sort out these possibilities.

```
         **: * .***.:*****.:*.:*******:*:*****:**** ***:*:*****:*                     .**.:*:*****. ..:. . :** ** **..
Sim-51  MKLFTSVCAIFVCIALFGATAEAGPFFFGAIRRFVEPFVPKVNATNSGTSFQLPFGLGGGSTESNGDEGASGYASGPDGSVGASGSGSGASGSGSANGSGASGSGSAGADANADSSQGNDYDSDYSQ
Sim-52  MKLFTSVCAIFVCIALFGATAEAGPFFFGAIRRFVEPFVPKVNATNSGTSIQLPFGLGGGSTESNGDGGASGYASGPDGSVGASGSGSGASGSGSANGSGASGSGSAGADANADSSQGNDYDSDYSQ
Sim-53  MKLFTSVCAIFVCIALFGATAEAGPFFFGAIRRFVEPFVPKVNATNSGTSFQLPFGLGGGSTESNGDGGASGYASGPDGSVGASGSGSGASGSGSANGSGASGSGSAGADANADSSQGNDYDSDYSQ
Sim-59  MKLFTTVCAIVVCIALFGATAEAGPFFFGAIRRFVEPFVPKVNATNSGTSFQLPFGLGGGSTESNGDGGASGYASGPNGSVGASGSGSGASGSGSANGSGASGSGSAGADANADSSQGNDYDSDYSQ
Sim-58  MKLFTTVCAIVVCIALFGATAEAGPFFFGAIRRFVEPFVPKVNATNSGTSIQLPFGLGGGSTESNGDGGASGYASGPDGSVGASGSGSGASGSGSANGSGASGSGSAGADANADSSQGNDYDSDYSQ
SIMVI_197260714  MKFATIVCAIILCIALFSTTVDAGPFFFGTIKRFVEPFIPKVNADNSGSSIQLPFGFG-----------------------------NASAAADSGASADGDSSQTNNSS---DYTDYEN
SIMVI_197260712  MKFATIACAIVLCIALFSTTVDAGPFFFGTIKRFVEPFIPKVNADNSGSSIQLPFGFG-----------------------------NASAAADSGASAAGDSSQTTNSS---DYTDYES
```

FIGURE 10. The Sv 7kDa family of black flies. The *Simulium nigrimanum* proteins are indicated by Sim-XX where XX is the number indicated in supplemental file S2. The remaining sequences are from *Simulium vittatum*, with indicated National Center for Biotechnology Information (NCBI) protein accession number. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. For more details, see text. This figure appears in color at www.ajtmh.org.

8–10 CysW family. Eight deducted proteins from the *S. nigrimanum* sialotranscriptome have similar sizes and have a common framework of tryptophans cysteines and other aliphatic residues (Figure 12). Two subfamilies are clearly distinguished, one containing 10 conserved cysteines and five conserved tryptophans, the other containing eight and six conserved C and W residues, respectively. They produce only low score matches to proteins in the NR or Swissprot databases. To obtain further insight on this family, the two best matches from the NR database were used to build a Psiblast search model with each member of the 8–10 Cys family. Interestingly, the 8 Cys member Sim-159 retrieves, after one iteration of Psiblast, proteins annotated as nicotinic acetylcholine receptor, with small e-values ($5e^{-43}$), whereas Sim-145 processed similarly retrieves proteins annotated as junctional adhesion molecules, with similar small e-values. Interestingly, whereas autoimmunity to desmoglein is proposed as the cause of pemphigus, other targets of autoimmunity have been considered, including the nicotinic acetylcholine receptor,[103,104] while junctional molecules are profoundly involved in skin and vascular epithelium adhesion.[105] Although these comparisons may not be robust, they provide for candidate proteins that could be implicated in the etiology of pemphigus foliaceus.

Additional orphan S. nigrimanum *proteins.* We additionally identified seven coding sequences in the *S. nigrimanum* sialotranscriptome that code for putative secreted proteins, based on 1–4 EST each. These are annotated as Orphan *S. nigrimanum* proteins in supplemental file S2. We additionally report four additional similarly orphan proteins, which are show alleles, based on multiple ESTs. These are named the acidic 28 kDa, the basic 28 kDa, the Sn basic 4.4 kDa, and the Sn basic 17 kDa families. They have no significant matches to any database queried in Supplemental Table S2.

Comparison of the *S. nigrimanum* sialome to desmogleins. To help identify possible candidate proteins that could induce anti-desmoglein immunity, we obtained a subset of the NR database that contained the word desmoglein in its fasta definition line and used the tool BLASTP to search the sialome against this database (using the switches –W2

and –FF, meaning word size for search = 2 and filter of low complexity turned off). The results are shown in column AV of Supplemental Table S2. The protein family named Collagen-like produced matches with e-values lower than 0.0001, mainly caused by the Gly repeats in the protein. The Simulium mucin family also produced small e-values in the comparisons, again caused by the glycine repeats. To the extent this family is glycosylated (observe the serine and threonines in between the glycines), these modifications should be taken into account. The Sv7.0 family also shows similarities to desmoglein, caused by glycine and also aspartyl repeats. It has been pointed out previously the possible similarities of the 8 Cys family to the nicotinic acetylcholine receptor and to junctional adhesion molecule families, which may be implicated in pemphigus pathology.

Housekeeping proteins. Supplemental Table S2 presents sequence information on 99 proteins classified as housekeeping.

CONCLUSION

Analysis of the sialome of *S. nigrimanum*, the second done for this family of blood-feeding flies, uncovers both the common and divergent evolutionary pathways taken in producing today's salivary "magic potion" of such arthropods. The previously reported *S. vittatum* sialotranscriptome revealed protein families unique to *Simulium*, those unique to blood-feeding Diptera, and ubiquitous protein families.[15] The *S. nigrimanum* sialotranscriptome indeed has these three classes of salivary proteins, including 13 protein families unique to *Simulium*, but additionally uncovered 13 novel protein classes, including six that are relatively well expressed. It is interesting to notice that most of the *S. nigrimanum* salivary proteins have only 30–60% identity to their *S. vittatum* best match, whereas the proteins from the housekeeping class have over 95% identity, thus showing the fast divergence of the salivary proteins.

The finding of many new protein families in *S. nigrimanum* as compared with *S. vittatum* may be related to the anthropophilic behavior of *S. nigrimanum* as compared with *S. vittatum*, as well as the fact that *S. vittatum* is autogenous. The
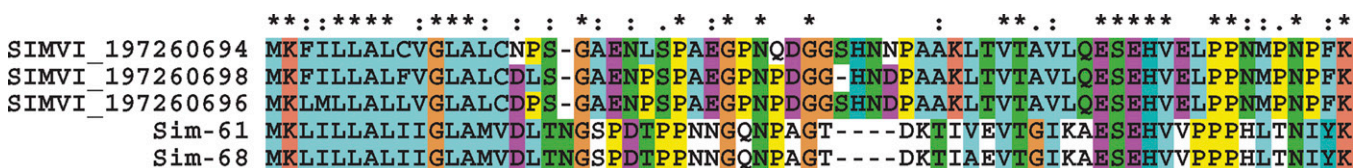


```
                 **::**** :***: : : *: : .* :* * *       :  **.: ***** **::.* :*
SIMVI_197260694  MKFILLALCVGLALCNPS-GAENLSPAEGPNQDGGSHNNPAAKLTVTAVLQESEHVELPPNMPNPFK
SIMVI_197260698  MKFILLALFVGLALCDLS-GAENPSPAEGPNPDGG-HNDPAAKLTVTAVLQESEHVELPPNMPNPFK
SIMVI_197260696  MKLMLLALLVGLALCDPS-GAENPSPAEGPNPDGGSHNDPAAKLTVTAVLQESEHVELPPNMPNPFK
       Sim-61    MKLILLALIIGLAMVDLTNGSPDTPPNNGQNPAGT----DKTIVEVTGIKAESEHVVPPPHLTNIYK
       Sim-68    MKLILLALIIGLAMVDLTNGSPDTPPNNGQNPAGT----DKTIAEVTGIKAESEHVVPPPHLTNIYK
```

FIGURE 11. The Sv 4.8kDa family of black flies. The *Simulium nigrimanum* proteins are indicated by Sim-XX where XX is the number indicated in supplemental file S2. The remaining sequences are from *Simulium vittatum*, with indicated National Center for Biotechnology Information (NCBI) protein accession number. The symbols above the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. For more details, see text. This figure appears in color at www.ajtmh.org.
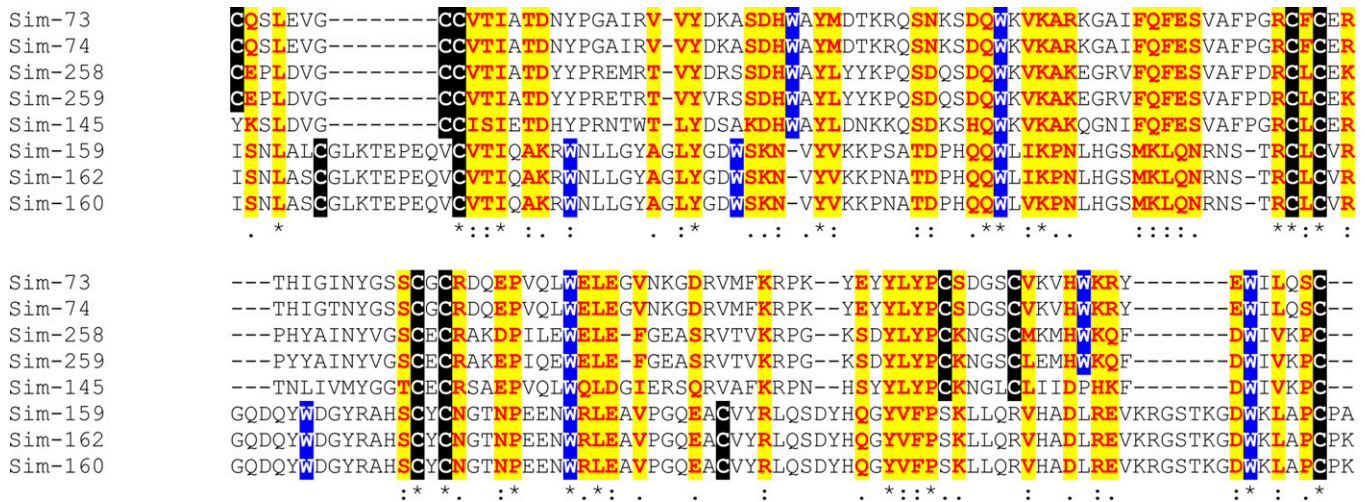
FIGURE 12. The 8-10CysW family of *Simulium nigrimanum*. The signal peptide region is not shown. The symbols below the alignment indicate: (*) identical sites; (:) conserved sites; (.) less conserved sites. Cysteines, tryptophanes, and conserved residues are marked with different backgrounds Protein sequences were aligned by the Clustal program39. For more details, see text. This figure appears in color at www.ajtmh.org.

dependence of a mammalian host for reproduction could have increased the pressure for the best possible salivary potion in *S. nigrimanum* as opposed to the more relaxed state of an autogenous species.

It is also to be noted that this sialotranscriptome generated many EST clusters coding for very similar proteins, having only a few AA changes, suggesting the existence of alleles or gene duplications creating very closely related genes, as indicated throughout the Results section. This is in contrast to many sialotranscriptomes performed so far, which derived from colonized material that has suffered bottlenecks and should be more genetically homogeneous. This indicates that salivary gland genes may have a tendency to be polymorphic and might be good population markers; it may also relate to the patchy distribution of Fogo Selvagem to the extent that it may be promoted by specific alleles producing cross-reactivity with self antigens.

Authors' addresses: José M. C. Ribeiro, Jesus G. Valenzuela, Van M. Pham, and Lindsay Kleeman, National Institute of Allergy and Infectious Diseases, Laboratory of Malaria and Vector Research, Rockville, MD, E-mails: jribeiro@niaid.nih.gov, vpham@niaid.nih.gov, jvalenzuela@niaid.nih.gov, and kleemanli@niaid.nih.gov. Kent D. Barbian and Amanda J. Favreau, Genomics Unit, Research Technologies Section, Rocky Mountain Laboratories, Hamilton, MT, E-mails: KBarbian@niaid.nih.gov and favreaua@niaid.nih.gov. Donald P. Eaton, Wildlife Conservation Society, Campo Grande, MS, Brazil, E-mail: ksadeaton@yahoo.com. Valeria Aoki and Evandro A. Rivitti, Department of Dermatology, Universidade de São Paulo, Brazil, E-mails: valaoki@hotmail.com and elimaria@hcnet.usp.br. Gunter Hans-Filho, Department of Dermatology, Universidade Federal de Mato Grosso do Sul, Campo Grande, Brazil, E-mail: ghansfilho@hotmail.com. Luis A. Diaz, Department of Dermatology, University of North Carolina at Chapel Hill, Chapel Hill, NC, E-mail: ldiaz@med.unc.edu.

REFERENCES

1. Culton DA, Qian Y, Li N, Rubenstein D, Aoki V, Filho GH, Rivitti EA, Diaz LA, 2008. Advances in pemphigus and its endemic pemphigus foliaceus (Fogo Selvagem) phenotype: a paradigm of human autoimmunity. *J Autoimmun 31:* 311–324.

2. Qian Y, Clarke SH, Aoki V, Hans-Filho G, Rivitti EA, Diaz LA, 2009. Antigen selection of anti-DSG1 autoantibodies during and before the onset of endemic pemphigus foliaceus. *J Invest Dermatol 129:* 2823–2834.

3. Rock B, Labib RS, Diaz LA, 1990. Monovalent Fab' immunoglobulin fragments from endemic pemphigus foliaceus autoantibodies reproduce the human disease in neonatal Balb/c mice. *J Clin Invest 85:* 296–299.

4. Hilario-Vargas J, Dasher DA, Li N, Aoki V, Hans-Filho G, dos Santos V, Qaqish BF, Rivitti EA, Diaz LA, 2006. Prevalence of anti-desmoglein-3 antibodies in endemic regions of Fogo selvagem in Brazil. *J Invest Dermatol 126:* 2044–2048.

5. Eaton DP, Diaz LA, Hans-Filho G, Santos VD, Aoki V, Friedman H, Rivitti EA, Sampaio SA, Gottlieb MS, Giudice GJ, Lopez A, Cupp EW, 1998. Comparison of black fly species (Diptera: Simuliidae) on an Amerindian reservation with a high prevalence of fogo selvagem to neighboring disease-free sites in the State of Mato Grosso do Sul, Brazil. The Cooperative Group on Fogo Selvagem Research. *J Med Entomol 35:* 120–131.

6. Ribeiro JM, Francischetti IM, 2003. Role of arthropod saliva in blood feeding: sialome and post-sialome perspectives. *Annu Rev Entomol 48:* 73–88.

7. Ribeiro JM, 1995. Blood-feeding arthropods: live syringes or invertebrate pharmacologists? *Infect Agents Dis 4:* 143–152.

8. Ribeiro JMC, Arcà B, 2009. From sialomes to the sialoverse: an insight into the salivary potion of blood feeding insects. *Adv Insect Physiol 37:* 59–118.

9. Calvo E, Pham VM, Marinotti O, Andersen JF, Ribeiro JM, 2009. The salivary gland transcriptome of the neotropical malaria

vector *Anopheles darlingi* reveals accelerated evolution of genes relevant to hematophagy. *BMC Genomics 10:* 57.

10. Ribeiro JM, Arca B, Lombardo F, Calvo E, Phan VM, Chandra PK, Wikel SK, 2007. An annotated catalogue of salivary gland transcripts in the adult female mosquito, *Aedes aegypti. BMC Genomics 8:* 6.

11. Arca B, Lombardo F, Francischetti IM, Pham VM, Mestres-Simon M, Andersen JF, Ribeiro JM, 2007. An insight into the sialome of the adult female mosquito *Aedes albopictus. Insect Biochem Mol Biol 37:* 107–127.

12. Arca B, Lombardo F, Valenzuela JG, Francischetti IM, Marinotti O, Coluzzi M, Ribeiro JM, 2005. An updated catalogue of salivary gland transcripts in the adult female mosquito, *Anopheles gambiae. J Exp Biol 208:* 3971–3986.

13. Ribeiro JM, Charlab R, Pham VM, Garfield M, Valenzuela JG, 2004. An insight into the salivary transcriptome and proteome of the adult female mosquito *Culex pipiens quinquefasciatus. Insect Biochem Mol Biol 34:* 543–563.

14. Valenzuela JG, Francischetti IM, Pham VM, Garfield MK, Ribeiro JM, 2003. Exploring the salivary gland transcriptome and proteome of the *Anopheles stephensi* mosquito. *Insect Biochem Mol Biol 33:* 717–732.

15. Andersen JF, Pham VM, Meng Z, Champagne DE, Ribeiro JM, 2009. Insight into the sialome of the Black Fly, *Simulium vittatum. J Proteome Res 8:* 1474–1488.

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

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

18. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG, 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res 25:* 4876–4882.

19. Kumar S, Tamura K, Nei M, 2004. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform 5:* 150–163.

20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G, 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet 25:* 25–29.

21. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL, 2000. The Pfam protein families database. *Nucleic Acids Res 28:* 263–266.

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

23. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA, 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics 4:* 41.

24. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH, 2002. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res 30:* 281–283.

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

26. Julenius K, Molgaard A, Gupta R, Brunak S, 2005. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology 15:* 153–164.

27. Friedman R, Hughes AL, 2005. Codon volatility as an indicator of positive selection: data from eukaryotic genome comparisons. *Mol Biol Evol 22:* 542–546.

28. Silva JC, Loreto EL, Clark JB, 2004. Factors that affect the horizontal transfer of transposable elements. *Curr Issues Mol Biol 6:* 57–71.

29. Govan VA, Leat N, Allsopp M, Davison S, 2000. Analysis of the complete genome sequence of acute bee paralysis virus shows that it belongs to the novel group of insect-infecting RNA viruses. *Virology 277:* 457–463.

30. Koonin EV, Wolf YI, Nagasaki K, Dolja VV, 2008. The Big Bang of picorna-like virus evolution antedates the radiation of eukaryotic supergroups. *Nat Rev Microbiol 6:* 925–939.

31. Stanway G, Brown F, Christian P, Hovi T, Hyypiä T, King AMQ, Knowles NJ, Lemon SM, Minor PD, Pallansch MA, Palmenberg AC, Skern T, 2005. Family *Picornaviridae. Virus Taxonomy.* Eighth Report of the International Committee on Taxonomy of Viruses. London: Elsevier/Academic Press, 757–778.

32. Tate J, Liljas L, Scotti P, Christian P, Lin T, Johnson JE, 1999. The crystal structure of cricket paralysis virus: the first view of a new virus family. *Nat Struct Biol 6:* 765–774.

33. Francischetti IM, Valenzuela JG, Pham VM, Garfield MK, Ribeiro JM, 2002. Toward a catalog for the transcripts and proteins (sialome) from the salivary gland of the malaria vector *Anopheles gambiae. J Exp Biol 205:* 2429–2451.

34. Ribeiro JM, Andersen J, Silva-Neto MA, Pham VM, Garfield MK, Valenzuela JG, 2004. Exploring the sialome of the blood-sucking bug *Rhodnius prolixus. Insect Biochem Mol Biol 34:* 61–79.

35. Galperin MY, Koonin EV, 2004. "Conserved hypothetical" proteins: prioritization of targets for experimental study. *Nucleic Acids Res 32:* 5452–5463.

36. Kanost MR, Jiang H, Yu XQ, 2004. Innate immune responses of a lepidopteran insect, Manduca sexta. *Immunol Rev 198:* 97–105.

37. Soderhall K, Cerenius L, 1998. Role of the prophenoloxidase-activating system in invertebrate immunity. *Curr Opin Immunol 10:* 23–28.

38. Volfova V, Hostomska J, Cerny M, Votypka J, Volf P, 2008. Hyaluronidase of bloodsucking insects and its enhancing effect on leishmania infection in mice. *PLoS Negl Trop Dis 2:* e294.

39. Wilson AD, Heesom KJ, Mawby WJ, Mellor PS, Russell CL, 2008. Identification of abundant proteins and potential allergens in *Culicoides nubeculosus* salivary glands. *Vet Immunol Immunopathol 122:* 94–103.

40. Ribeiro JM, Charlab R, Rowton ED, Cupp EW, 2000. *Simulium vittatum* (Diptera: Simuliidae) and *Lutzomyia longipalpis* (Diptera: Psychodidae) salivary gland hyaluronidase activity. *J Med Entomol 37:* 743–747.

41. Wang X, Ribeiro JM, Broce AB, Wilkerson MJ, Kanost MR, 2009. An insight into the transcriptome and proteome of the salivary gland of the stable fly, *Stomoxys calcitrans. Insect Biochem Mol Biol 39:* 607–614.

42. Calvo E, Ribeiro JM, 2006. A novel secreted endonuclease from *Culex quinquefasciatus* salivary glands. *J Exp Biol 209:* 2651–2659.

43. Valenzuela JG, Garfield M, Rowton ED, Pham VM, 2004. Identification of the most abundant secreted proteins from the salivary glands of the sand fly *Lutzomyia longipalpis*, vector of *Leishmania chagasi. J Exp Biol 207:* 3717–3729.

44. Ribeiro JM, 1987. Role of arthropod saliva in blood feeding. *Annu Rev Entomol 32:* 463–478.

45. Sun D, McNicol A, James AA, Peng Z, 2006. Expression of functional recombinant mosquito salivary apyrase: a potential therapeutic platelet aggregation inhibitor. *Platelets 17:* 178–184.

46. Champagne DE, Smartt CT, Ribeiro JM, James AA, 1995. The salivary gland-specific apyrase of the mosquito *Aedes aegypti* is a member of the 5′-nucleotidase family. *Proc Natl Acad Sci USA 92:* 694–698.

47. Faudry E, Santana JM, Ebel C, Vernet T, Teixeira AR, 2006. Salivary apyrases of *Triatoma infestans* are assembled into homo-oligomers. *Biochem J 396:* 509–515.

48. Faudry E, Lozzi SP, Santana JM, D'Souza-Ault M, Kieffer S, Felix CR, Ricart CA, Sousa MV, Vernet T, Teixeira AR, 2004. *Triatoma infestans* apyrases belong to the 5′-nucleotidase family. *J Biol Chem 279:* 19607–19613.

49. Valenzuela JG, Belkaid Y, Rowton E, Ribeiro JM, 2001. The salivary apyrase of the blood-sucking sand fly *Phlebotomus papatasi* belongs to the novel *Cimex* family of apyrases. *J Exp Biol 204:* 229–237.

50. Valenzuela JG, Charlab R, Galperin MY, Ribeiro JM, 1998. Purification, cloning, and expression of an apyrase from the bed bug *Cimex lectularius*. A new type of nucleotide-binding enzyme. *J Biol Chem 273:* 30583–30590.

51. Andersen JF, Hinnebusch BJ, Lucas DA, Conrads TP, Veenstra TD, Pham VM, Ribeiro JM, 2007. An insight into the sialome of the oriental rat flea, *Xenopsylla cheopis* (Rots). *BMC Genomics 8:* 102.

52. Ogata S, Hayashi Y, Misumi Y, Ikehara Y, 1990. Membrane-anchoring domain of rat liver 5′-nucleotidase: identification of the COOH-terminal serine-523 covalently attached with a glycolipid. *Biochemistry 29:* 7923–7927.

53. Misumi Y, Ogata S, Ohkubo K, Hirose S, Ikehara Y, 1990. Primary structure of human placental 5′-nucleotidase and identification of the glycolipid anchor in the mature form. *Eur J Biochem 191:* 563–569.

54. Cupp MS, Cupp EW, Ochoa AJ, Moulton JK, 1995. Salivary apyrase in New World blackflies (Diptera: Simuliidae) and its relationship to onchocerciasis vector status. *Med Vet Entomol 9:* 325–330.

55. Baskova IP, Nikonov GI, 1991. Destabilase, the novel epsilon-(gamma-Glu)-Lys isopeptidase with thrombolytic activity. *Blood Coagul Fibrinolysis 2:* 167–172.

56. Zavalova LL, Artamonova II, Berezhnoy SN, Tagaev AA, Baskova IP, Andersen J, Roepstorff P, Egorov Ts A, 2003. Multiple forms of medicinal leech destabilase-lysozyme. *Biochem Biophys Res Commun 306:* 318–323.

57. Zavalova LL, Baskova IP, Lukyanov SA, Sass AV, Snezhkov EV, Akopov SB, Artamonova II, Archipova VS, Nesmeyanov VA, Kozlov DG, Benevolensky SV, Kiseleva VI, Poverenny AM, Sverdlov ED, 2000. Destabilase from the medicinal leech is a representative of a novel family of lysozymes. *Biochim Biophys Acta 1478:* 69–77.

58. Vija H, Samel M, Siigur E, Aaspollu A, Tonismagi K, Trummal K, Subbi J, Siigur J, 2009. VGD and MLD-motifs containing heterodimeric disintegrin viplebedin-2 from Vipera lebetina snake venom. Purification and cDNA cloning. *Comp Biochem Physiol B Biochem Mol Biol 153:* 253–260.

59. Calvete JJ, Moreno-Murciano MP, Theakston RD, Kisiel DG, Marcinkiewicz C, 2003. Snake venom disintegrins: novel dimeric disintegrins and structural diversification by disulphide bond engineering. *Biochem J 372:* 725–734.

60. Ascenzi P, Bocedi A, Bolognesi M, Spallarossa A, Coletta M, De Cristofaro R, Menegatti E, 2003. The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein. *Curr Protein Pept Sci 4:* 231–251.

61. Salier JP, 1990. Inter-alpha-trypsin inhibitor: emergence of a family within the Kunitz-type protease inhibitor superfamily. *Trends Biochem Sci 15:* 435–439.

62. Paesen GC, Siebold C, Dallas ML, Peers C, Harlos K, Nuttall PA, Nunn MA, Stuart DI, Esnouf RM, 2009. An ion-channel modulator from the saliva of the brown ear tick has a highly modified Kunitz/BPTI structure. *J Mol Biol 389:* 734–747.

63. Castaneda O, Harvey AL, 2009. Discovery and characterization of cnidarian peptide toxins that affect neuronal potassium ion channels. *Toxicon 54:* 1119–1124.

64. Dy CY, Buczek P, Imperial JS, Bulaj G, Horvath MP, 2006. Structure of conkunitzin-S1, a neurotoxin and Kunitz-fold disulfide variant from cone snail. *Acta Crystallogr D Biol Crystallogr 62:* 980–990.

65. Lucchesi KJ, Moczydlowski E, 1991. On the interaction of bovine pancreatic trypsin inhibitor with maxi Ca(2+)-activated K+ channels. A model system for analysis of peptide-induced subconductance states. *J Gen Physiol 97:* 1295–1319.

66. Mans BJ, Neitz AW, 2004. Adaptation of ticks to a blood-feeding environment: evolution from a functional perspective. *Insect Biochem Mol Biol 34:* 1–17.

67. Steen NA, Barker SC, Alewood PF, 2006. Proteins in the saliva of the Ixodida (ticks): pharmacological features and biological significance. *Toxicon 47:* 1–20.

68. Francischetti IM, Valenzuela JG, Andersen JF, Mather TN, Ribeiro JM, 2002. Ixolaris, a novel recombinant tissue factor pathway inhibitor (TFPI) from the salivary gland of the tick, *Ixodes scapularis:* identification of factor X and factor Xa as scaffolds for the inhibition of factor VIIa/tissue factor complex. *Blood 99:* 3602–3612.

69. Stark KR, James AA, 1998. Isolation and characterization of the gene encoding a novel factor Xa-directed anticoagulant from the yellow fever mosquito, *Aedes aegypti. J Biol Chem 273:* 20802–20809.

70. Valenzuela JG, Francischetti IM, Ribeiro JM, 1999. Purification, cloning, and synthesis of a novel salivary anti-thrombin from the mosquito *Anopheles albimanus. Biochemistry 38:* 11209–11215.

71. Francischetti IM, Valenzuela JG, Ribeiro JM, 1999. Anophelin: kinetics and mechanism of thrombin inhibition. *Biochemistry 38:* 16678–16685.

72. Campbell CL, Vandyke KA, Letchworth GJ, Drolet BS, Hanekamp T, Wilson WC, 2005. Midgut and salivary gland transcriptomes of the arbovirus vector *Culicoides sonorensis* (Diptera: Ceratopogonidae). *Insect Mol Biol 14:* 121–136.

73. Andersen JF, Gudderra NP, Francischetti IM, Valenzuela JG, Ribeiro JM, 2004. Recognition of anionic phospholipid membranes by an antihemostatic protein from a blood-feeding insect. *Biochemistry 43:* 6987–6994.

74. Buser CA, Kim J, McLaughlin S, Peitzsch RM, 1995. Does the binding of clusters of basic residues to acidic lipids induce domain formation in membranes? *Mol Membr Biol 12:* 69–75.

75. Kirszberg C, Lima LG, Da Silva de Oliveira A, Pickering W, Gray E, Barrowcliffe TW, Rumjanek VM, Monteiro RQ, 2009. Simultaneous tissue factor expression and phosphatidylserine exposure account for the highly procoagulant pattern of melanoma cell lines. *Melanoma Res 19:* 301–308.

76. Megraw T, Kaufman TC, Kovalick GE, 1998. Sequence and expression of *Drosophila* Antigen 5-related 2, a new member of the CAP gene family. *Gene 222:* 297–304.

77. Milne TJ, Abbenante G, Tyndall JD, Halliday J, Lewis RJ, 2003. Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily. *J Biol Chem 278:* 31105–31110.

78. Yamazaki Y, Hyodo F, Morita T, 2003. Wide distribution of cysteine-rich secretory proteins in snake venoms: isolation and cloning of novel snake venom cysteine-rich secretory proteins. *Arch Biochem Biophys 412:* 133–141.

79. Yamazaki Y, Morita T, 2004. Structure and function of snake venom cysteine-rich secretory proteins. *Toxicon 44:* 227–231.

80. Mochca-Morales J, Martin BM, Possani LD, 1990. Isolation and characterization of helothermine, a novel toxin from *Heloderma horridum horridum* (Mexican beaded lizard) venom. *Toxicon 28:* 299–309.

81. Xu X, Yang H, Ma D, Wu J, Wang Y, Song Y, Wang X, Lu Y, Yang J, Lai R, 2008. Toward an understanding of the molecular mechanism for successful blood feeding by coupling proteomics analysis with pharmacological testing of horsefly salivary glands. *Mol Cell Proteomics 7:* 582–590.

82. Scarborough RM, Naughton MA, Teng W, Rose JW, Phillips DR, Nannizzi L, Arfsten A, Campbell AM, Charo IF, 1993. Design of potent and specific integrin antagonists. Peptide antagonists with high specificity for glycoprotein IIb-IIIa. *J Biol Chem 268:* 1066–1073.

83. Ameri M, Wang X, Wilkerson MJ, Kanost MR, Broce AB, 2008. An immunoglobulin binding protein (antigen 5) of the stable fly (Diptera: Muscidae) salivary gland stimulates bovine immune responses. *J Med Entomol 45:* 94–101.

84. Valenzuela JG, Charlab R, Gonzalez EC, de Miranda-Santos IK, Marinotti O, Francischetti IM, Ribeiro JM, 2002. The D7 family of salivary proteins in blood sucking diptera. *Insect Mol Biol 11:* 149–155.

85. Hekmat-Scafe DS, Dorit RL, Carlson JR, 2000. Molecular evolution of odorant-binding protein genes OS-E and OS-F in *Drosophila. Genetics 155:* 117–127.

86. Mans BJ, Calvo E, Ribeiro JM, Andersen JF, 2007. The crystal structure of D7r4, a salivary biogenic amine-binding protein from the malaria mosquito *Anopheles gambiae. J Biol Chem 282:* 36626–36633.

87. Valenzuela JG, Charlab R, Gonzalez EC, Miranda-Santos IKF, Marinotti O, Francischetti IM, Ribeiro JMC, 2002. The D7 family of salivary proteins in blood sucking Diptera. *Insect Mol Biol 11:* 149–155.

88. Calvo E, Mans BJ, Andersen JF, Ribeiro JM, 2006. Function and evolution of a mosquito salivary protein family. *J Biol Chem 281:* 1935–1942.

89. Calvo E, Mans BJ, Ribeiro JM, Andersen JF, 2009. Multifunctionality and mechanism of ligand binding in a mosquito antiinflammatory protein. *Proc Natl Acad Sci USA 106:* 3728–3733.

90. Cupp MS, Cupp EW, 2000. *Antithrombin Protein and DNA Sequences from Black Fly.* USA.

91. Stevenson KJ, Poller L, 1982. The procoagulant activity of partial thromboplastin extracts: the role of phosphatidyl serine. *Thromb Res 26:* 341–350.

92. Simons FE, Peng Z, 2001. Mosquito allergy: recombinant mosquito salivary antigens for new diagnostic tests. *Int Arch Allergy Immunol 124:* 403–405.

93. Calvo E, Tokumasu F, Marinotti O, Villeval JL, Ribeiro JM, Francischetti IM, 2007. Aegyptin, a novel mosquito salivary gland protein, specifically binds to collagen and prevents its interaction with platelet glycoprotein VI, integrin alpha-2beta1, and von Willebrand factor. *J Biol Chem 282:* 26928–26938.

94. Yoshida S, Sudo T, Niimi M, Tao L, Sun B, Kambayashi J, Watanabe H, Luo E, Matsuoka H, 2008. Inhibition of collagen-induced platelet aggregation by anopheline antiplatelet protein, a saliva protein from a malaria vector mosquito. *Blood 111:* 2007–2014.

95. Grimaldi D, Engel M, 2005. *Evolution of the Insects*. New York: Cambridge University Press.

96. Cupp MS, Ribeiro JM, Cupp EW, 1994. Vasodilative activity in black fly salivary glands. *Am J Trop Med Hyg 50:* 241–246.

97. Cupp MS, Ribeiro JM, Champagne DE, Cupp EW, 1998. Analyses of cDNA and recombinant protein for a potent vasoactive protein in saliva of a blood-feeding black fly, *Simulium vittatum*. *J Exp Biol 201:* 1553–1561.

98. Pinnell SR, 1985. Regulation of collagen biosynthesis by ascorbic acid: a review. *Yale J Biol Med 58:* 553–559.

99. Rydengard V, Andersson Nordahl E, Schmidtchen A, 2006. Zinc potentiates the antibacterial effects of histidine-rich peptides against *Enterococcus faecalis*. *FEBS J 273:* 2399–2406.

100. Nishikawa M, Ogawa K, 2004. Antimicrobial activity of a chelatable poly(arginyl-histidine) produced by the ergot fungus *Verticillium kibiense*. *Antimicrob Agents Chemother 48:* 229–235.

101. Loomans HJ, Hahn BL, Li QQ, Phadnis SH, Sohnle PG, 1998. Histidine-based zinc-binding sequences and the antimicrobial activity of calprotectin. *J Infect Dis 177:* 812–814.

102. Tanaka H, Chiba H, Inokoshi J, Kuno A, Sugai T, Takahashi A, Ito Y, Tsunoda M, Suzuki K, Takenaka A, Sekiguchi T, Umeyama H, Hirabayashi J, Omura S, 2009. Mechanism by which the lectin actinohivin blocks HIV infection of target cells. *Proc Natl Acad Sci USA 106:* 15633–15638.

103. Femiano F, 2007. Pemphigus vulgaris: recent advances in our understanding of its pathogenesis. *Minerva Stomatol 56:* 215–223.

104. Grando SA, 2006. Cholinergic control of epidermal cohesion. *Exp Dermatol 15:* 265–282.

105. Bazzoni G, 2003. The JAM family of junctional adhesion molecules. *Curr Opin Cell Biol 15:* 525–530.