Published in final edited form as:

Insect Biochem Mol Biol. 2015 October ; 65: 83–90. doi:10.1016/j.ibmb.2015.09.002.

An insight into the sialome of the horse fly, *Tabanus bromius*

José M.C. Ribeiro¹, Maria Kazimirova², Peter Takac², John F. Andersen¹, and Ivo M.B. Francischetti¹

¹National Institute of Allergy and Infectious Diseases, Laboratory of Malaria and Vector Research, 12735 Twinbrook Parkway, Maryland 20852, USA

²Institute of Zoology, Slovak Academy of Sciences, Dubravská cesta 9, SK-84506 Bratislava, Slovakia

Abstract

Blood feeding animals face their host's defenses against tissue injury and blood loss while attempting to feed. One adaptation to surmount these barriers involves the evolution of a salivary potion that disarms their host's inflammatory and anti-hemostatic processes. The composition of the peptide moiety of this potion, or sialome (from the Greek sialo=saliva), can be deduced in part by proper interpretation of the blood feeder's sialotranscriptome. In this work we disclose the sialome of the blood feeding adult female *Tabanus bromius*. Following assembly of over 75 million Illumina reads (101 nt long) 16,683 contigs were obtained from which 4,078 coding sequences were extracted. From these, 320 were assigned as coding for putative secreted proteins. These 320 contigs mapped 85% of the reads. The antigen-5 proteins family was studied in detail, indicating three *Tabanus* specific clades with and without disintegrin domains, as well as with and without leukotriene binding domains. Defensins were also detailed; a clade of salivary tabanid peptides was found lacking the propeptide domain ending in the KR dipeptide signaling furin cleavage. Novel protein families were also disclosed. Viral transcripts were identified closely matching the Kotonkan virus capsid proteins. Full length Mariner transposases were also identified. A total of 3,043 coding sequences and their protein products were deposited in Genbank. Hyperlinked excel spreadsheets containing the coding sequences and their annotation are available at http://exon.niaid.nih.gov/transcriptome/T_bromius/Tbromius-web.xlsx (hyperlinked excel spreadsheet, 11 MB) and http://exon.niaid.nih.gov/transcriptome/T_bromius/Tbromius-SA.zip (Standalone excel with all local links, 360 MB). These sequences provide for a platform from which further proteomic studies may be designed to identify salivary proteins from *T. bromius* that are of pharmacological interest or used as immunological markers of host exposure.

Corresponding author: José M.C. Ribeiro, National Institute of Allergy and Infectious Diseases, Laboratory of Malaria and Vector Research, 12735 Twinbrook Parkway, Maryland 20852, USA, jribeiro@nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Because J.M.C.R., J.F.A. and I.M.B.F. are government employees and this is a government work, the work is in the public domain in the United States. Notwithstanding any other agreements, the NIH reserves the right to provide the work to PubMedCentral for display and use by the public, and PubMedCentral may tag or modify the work consistent with its customary practices. You can establish rights outside of the U.S. subject to a government use license.

Keywords

medical entomology; veterinary entomology; tabanid; hematophagy; salivary glands; transcriptome

1. Introduction

Specialized mouthparts and a unique salivary composition are among the several adaptations observed in animals that evolve to a hematophagous habit. In the specific case of Tabanids, they blood-feed by lacerating their hosts' skin with their scissors-like mouthparts and sucking the blood from the ensuing hemorrhage, and as such are pool feeders, or telmophagic (Lavoipierre, 1965). Salivary gland products evolved to disarm their hosts' hemostasis, the mechanism preventing blood loss in vertebrates consisting of platelet aggregation, blood clotting and vasoconstriction. Additionally, salivary compounds also have immunoregulatory and antimicrobial functions, and in some insects they may assist sugar feeding (Ribeiro, 1995; Ribeiro and Arca, 2009). Adult tabanids take sugar meals, and only the females take additional blood meals, important for egg production in non-autogenous species (Friend and Stoffolano, 1991; Taylor and Smith, 1989). Salivary gland homogenates of tabanids have anti-clotting agents (Kazimirova et al., 2001; Kazimirova et al., 2002; Xu et al., 2008), a Kazal-type vasodilator (Takac et al., 2006; Xu et al., 2008; Zhang et al., 2014) and potent anti-platelet agents in the form of a modified protein of the antigen 5 family that recruited an RGD domain and additionally binds the pro-inflammatory cysteinyl leukotrienes (Ma et al., 2011b; Xu et al., 2012). A novel salivary protein from *Hybomitra atriperoides* named immunoregulin inhibits the secretion of interferon-gamma and monocyte chemoattractant protein (MCP-1) and increase IL-10 secretion in rat splenocytes activated by lipopolysaccharide (Yan et al., 2008; Zhao et al., 2009). Salivary enzymes associated with blood feeding have also been found in tabanids, including apyrase (An et al., 2011; Reddy et al., 2000), hyaluronidase (Ma et al., 2011a) and fibrinogenolytic serine proteases (Xu et al., 2008). A monumental sialotranscriptome associated with proteomics and bioassay screening was performed for the Asian *Tabanus yao*, yielded much biological information and the deposition of 53 protein sequences to GenBank (Ma et al., 2009; Xu et al., 2008).

In the present work we disclose the sialotranscriptome of the European species *Tabanus bromius*, a nuisance species of veterinary importance as a vector of *Trypanosoma theileri* (Bose et al., 1987; Dirie et al., 1990) and a potential vector of *Mycoplasma* infections to cattle (Hornok et al., 2011). An Illumina library from the salivary glands of adult female flies was “de novo” assembled, yielding 4,079 coding sequences from putative salivary and housekeeping functions, 3,043 of which appeared > 75% full length by blastp comparison to known proteins (Altschul et al., 1997), or were deemed to be novel full length proteins, and were accordingly submitted to GenBank. Novel protein families unique to *Tabanus* were found, as well as novel components of the tabanids sialome, such as endonucleases, phospholipases and dipeptidyl peptidases. This database will also serve as a reference library for future proteomic studies and to mine interesting pharmacologically active peptides from this fly.

2. Material and Methods

2.1. Insects

Adult female (100 individuals) *Tabanus bromius* were collected on July 2010 close to Vysoka pri Morave village, coordinates: 48°20'28" N; 16°55'08" E; altitude 140 m a.s.l., Slovakia. Their salivary glands were dissected and transferred to RNAlater.

2.2. RNA extraction, library preparation and sequencing

RNA preparation, library construction and sequencing were performed essentially as described previously (Ribeiro et al., 2013), and will be repeated here with small modifications for reference: SG were stored in RNAlater at 4°C for 48 h before being transferred to -70° C until RNA extraction. SG RNA was extracted and isolated using the Micro FastTrack mRNA isolation kit (Invitrogen, Grand Island, NY) per manufacturer's instructions. The integrity of the total RNA was checked on a Bioanalyser (Agilent Technologies, Santa Clara, CA). mRNA library construction and sequencing were done by the NIH Intramural Sequencing Center. The SG library was constructed using the TruSeq RNA sample prep kit, v. 2 (Illumina Inc., San Diego, CA). The resulting cDNA was fragmented using a Covaris E210 (Covaris, Woburn, MA). Library amplification was performed using eight cycles to minimize the risk of over-amplification. Sequencing was performed on a HiSeq 2000 (Illumina) with v. 3 flow cells and sequencing reagents. One lane of the HiSeq machine was used for this and four other libraries, distinguished by bar coding. Libraries from the tick *Hyalomma anatolicum excavatum*, the triatomine *Panstrongylus megistus* and the bat *Diphila ecaudata* were co-sequenced with *T. bromius* and we found that some cross-contamination of sequences occurred between the libraries. We are not sure how this contamination occurred as the RNA extractions were done at different times. It is possible that the bar coding primers were contaminated, or that the reads were misplaced at machine reading time. The degree of contamination varied from 0.1 to 1%. We raise this word of caution to other researchers when sharing lanes with unknown samples. A total of 75,723,876 reads were obtained for the *T. bromius* library, of 101 nt in length.

2.3. Bioinformatic tools used

This was done as previously described (Ribeiro et al., 2015) and is reproduced here with small changes. Raw data were processed using RTA 1.12.4.2 and CASAVA 1.8.2. Reads were trimmed of low quality regions (< 10), and only those with an average quality of 20 or more were used, comprising a total of 75,723,876 high-quality reads. Illumina primers were removed from the sequences following a parallel blastn of the reads against HiSeq TrueSeq adapters. Resulting reads were assembled with the ABySS software (Genome Sciences Centre, Vancouver, BC, Canada) (Birol et al., 2009; Simpson et al., 2009) using various kmer (k) values (every fifth from 21 to 91). Because the ABySS assembler tends to miss highly expressed transcripts (Zhao et al., 2011) the SOAPdenovo-Trans assembler (Luo et al., 2012) was also used, again with odd kmers from 21–91. The resulting assemblies were joined by an iterative BLAST and cap3 assembler (Karim et al., 2011). Sequence contamination between bar-coded libraries were identified and removed when their sequence identities were over 98% but their abundance of reads were > 10 fold between

libraries. Coding sequences (CDS) were extracted using an automated pipeline based on similarities to known proteins or by obtaining CDS containing a signal peptide (Nielsen et al., 1999). CDS and their protein sequences were mapped into a hyperlinked Excel spreadsheet (presented as Supplemental File 1). Signal peptide, transmembrane domains, furin cleavage sites, and mucin-type glycosylation were determined with software from the Center for Biological Sequence Analysis (Technical University of Denmark, Lyngby, Denmark) (Duckert et al., 2004; Julenius et al., 2005; Nielsen et al., 1999; Sonnhammer et al., 1998). Reads were mapped into the contigs using blastn (Altschul et al., 1997) with a word size of 25, masking homonucleotide decamers and allowing mapping to up to three different CDS if the BLAST results had the same score values. Mapping of the reads was also included in the Excel spreadsheet. FPKM values (Trapnell et al., 2012) for each coding sequence were also mapped to the spreadsheet. To compare relative expression of transcripts, we use the “expression index” defined as the number of reads mapped to a particular CDS multiplied by 100 and divided by the largest found number of reads mapped to a single CDS, which in the case of this transcriptome was a value of 6,059,575 mapped to a single tabimmunoregulin coding sequence. We also use “relative FPKM index” defined as above for “expression index”. The maximum value (100) was found to the same tabimmunoregulin CDS, which had an FPKM = 676,761. The “expression index” is to grams/liter as the “relative expression index” is to Moles/liter. Automated annotation of proteins was based on a vocabulary of nearly 290 words found in matches to various databases, including Swissprot, Gene Ontology, KOG, Pfam, and SMART, Refseq-invertebrates and a subset of the GenBank sequences containing diptera[organism] protein sequences, as well as the presence or not of signal peptides and transmembrane domains. Further manual annotation was done as required. Detailed bioinformatics analysis of our pipeline can be found in our previous publication (Karim et al., 2011). For determination of synonymous and non-synonymous sites within coding sequences, the tool BWA mem (Li and Durbin, 2010) was used to map the reads to the CDS into SAM files that were converted to BAM format, and sorted using Samtools (Li et al., 2009). The sequence alignment/map tools (samtools) package was used to do the mpileup of the reads (samtools mpileup), and the binary call format tools (bcftools) program from the same package was used to make the final vcf file containing the single-nucleotide polymorphic (SNP) sites, which were only taken if the site coverage was at least 100 (–D100), the quality was 13 or better and the SNP frequency was 5 or higher (default). Determination of whether the SNPs lead to a synonymous or non-synonymous codon change was achieved by a program written in Visual Basic by JMCR, the results of which are mapped into the Excel spreadsheet and color visualized in hyperlinked rtf files within Additional File 1. Sequence alignments were done with the ClustalX software package (Thompson et al., 1997). Phylogenetic analysis and statistical neighbor-joining bootstrap tests (1,000 iterations) of the phylogenies were done with the Mega package (Kumar et al., 2004). The antigen V proteins were modeled using homology methods on the Swiss-Model server (Schwede et al., 2003) based on the structure of tablysin-15 (PDB ID 3U3N).

2.4. Data access

The raw reads were deposited on the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI) under bioproject ID PRJNA283639 and run

SRR2017608. A total of 3,043 coding sequences and their protein products were deposited in the Transcriptome Shotgun Assembly portal of the NCBI to the DDBJ/EMBL/GenBank databases under the accession GDAI000000000. The version described in this paper is the first version, GDAI000000000. Hyperlinked excel spreadsheets containing the coding sequences and their annotation are available at http://exon.niaid.nih.gov/transcriptome/T_bromius/Tbromius-web.xlsx (hyperlinked excel spreadsheet, 11 MB) and http://exon.niaid.nih.gov/transcriptome/T_bromius/Tbromius-SA.zip (Standalone excel with all local links, 360 MB).

3. Results and Discussion

3.1. Overview of the sialotranscriptome of the adult female *Tabanus bromius*

Following assembly of 75,723,876 reads, a total of 16,683 assembled sequences were obtained that were equal or larger than 150 nt, having a median length of 1,338 nucleotides (nt). From these contigs, 4,078 coding sequences (CDS) were extracted. These coding sequences mapped 43,993,842 reads, or 58% of the total reads. Their median length was 1,257 nucleotides (nt) with 1,530 CDS being larger than 1,000 nt. These CDS were classified generally into five classes, namely “secreted” (S), “housekeeping” (H), “unknown” (U), “transposable elements” (TE) and “viral” (V) (Table 1). The S class had 320 assigned CDS, which mapped 85% of all reads, in accordance with the secretory nature of the organ. The H class produced 3,430 CDS, mapping 12% of the reads. TE's accounted for 2.6% of the CDS and 0.14% of the reads. Among these are included three transcripts coding for full length transposases of Mariner elements (TbTE-1037, TbTE-17225 and TbTE-16464), none having stop codons disrupting their coding sequences. Tb-4636 and Tb-16229 code for non long terminal repeats (NLTR) transposon proteins similar to LINE *Bilbo* transposon from *Drosophila* (Blesa and Martinez-Sebastian, 1997). Three putative viral transcripts were also found, mapping 0.07% of the reads. These include a transcript (Tb-1669) coding for a protein that has 51% amino acid (aa) similarity over a 227 aa stretch to the F protein of *Clanis bilineata nucleopolyhedrovirus* (Zhu et al., 2009) and expressed with a FPKM=2.2. It has a DUF3609 CDD domain found in eukaryotes and viruses. Two other transcripts present PFAM domains typical of Rhabdoviruses. TbSigP-7308 has the PFAM domain Rhabdo_glycop, coding for the Rhabdovirus spike glycoprotein and produces a best match to *Drosophila melanogaster sigmavirus* (Wilfert and Jiggins, 2014), characterized by 46% similarity over a 499 aa stretch. Tb-16195 has the domain PFAM domain Rhabdo_ncap, coding for the Rhabdovirus nucleocapsid protein, and producing matches to proteins from *Kotonkan virus*, having 50% similarity over 414 aa of its nucleocapsid major protein. This virus is associated with ephemeral cattle fever in Africa (Blasdell et al., 2012). Finally, 218 CDS were not able to be classified, representing 2% of the reads.

The housekeeping CDS were further classified by their function (Table 2), not surprisingly showing the category “protein synthesis” and “protein modification” to be the two most expressed and accumulating 33 and 11% of the reads of the H class, respectively. The class “unknown conserved” ranks third, with 8 % of the reads of the H class. This is similar to

other transcriptomes and reflects our ignorance of a significant proportion of proteins associated to basic cellular processes (Galperin and Koonin, 2004).

Further classification of the S class of transcripts reveals enzymes of the glucosidase and apyrase families to be abundantly expressed, as well as serine proteases and dipeptidyl peptidases. Members of the antigen 5 family are very abundant, accreting 54% of the reads of the S class, as are members of the unique tabanids family of tabimmunregulins, which accumulate 29% of the reads of the S class. Proteins with proteinase inhibitor domains of the Serpin, Kunitz and Kazal families were found, including members of the Vasotab family of tabanids vasodilators. Novel protein families, such as the Histidine-Proline rich family, accumulate 1.8% of the reads of the S class, as well as conserved secreted families of unknown function, including a family of proteins formerly found in sand fly midguts accumulating 6.6 % of the reads of the S class, and may function as an antimicrobial (Table 3).

3.2. Analysis of the putative secretory *Tabanus bromius* sialotranscriptome

3.2.1. Enzymes—Transcripts coding for amylases and glucosidases were found, including a highly expressed transcript coding for an alpha amylase with a relative read expression (RRE) of 6.3 and a relative FPKM expression (RFE) of 2.6. These enzymes most probably are associated to sugar feeding. Truncated transcripts code for enzymes of the CD73 ecto 5'-nucleotidase family that were previously described as salivary inhibitors of platelet aggregation from deer flies (Reddy et al., 2000), and also in *T. yao* (An et al., 2011). These apyrase-coding transcripts accumulate a RRE=8.5. An endonuclease-coding transcript (RRE=0.75) as well as one coding for a hyaluronidase (RRE=6.13) are abundantly expressed. These two enzymes are commonly found in pool feeders, such as sand flies, and might help to decrease skin viscosity and facilitate the diffusion of salivary constituents to the tissues (Ribeiro et al., 2010). Salivary endonucleases also affect the formation of neutrophil extracellular traps and enhance parasite transmission by sand flies (Chagas et al., 2014). A transcript coding for a heme-peroxidase (Tb-16332) is abundantly expressed (RRE=4.62); it is truncated at the 5' end thus it cannot be known whether it has a signal peptide indicative of secretion. However, its high RRE is indicative of a salivary role, as found uniquely in anopheline mosquito sialomes, where a salivary peroxidase functions as a catechol oxidase and serotonin peroxidase and act as a vasodilator (Ribeiro and Valenzuela, 1999; Ribeiro and Nussenzveig, 1993). Tb-16332 may be the ortholog of *T. yao* peroxidase activity previously reported, but not cloned (Ma et al., 2009). Phospholipase-coding transcripts of the A₂ family are abundant (RRE=2.17), and also found in mosquito and tick sialomes, but their roles in feeding are still uncertain. A chymotrypsin-like serine protease-coding transcript is abundantly expressed (RER=3.54), providing for 82% identity to tabserin from *T. yao* (Xu et al., 2008), and 55% identity to a flea salivary serine protease (gi|4530036) (Ribeiro et al., 2012). They might have a fibrinolytic function. A highly expressed dipeptidyl peptidase (RER=1.06) may function as a kininase, as found before in tick saliva (Ribeiro and Mather, 1998). Other peptidases are depicted in Supplemental File 1, but have low RRE's (< 0.1) and will not be further described.

3.2.2. Protease inhibitor domains—TbSigP-16659 has a Serpin Pfam domain and best matches a protein from *Glossina morsitans morsitans* (39% identity over 378 aa). It has a relatively low RRE (0.087). Kunitz and Kazal/Vasotab-coding transcripts have similar expression values. Some of the Kunitz domain transcripts are similar to anticlotting peptides previously reported from *T. yao* (Ma et al., 2009), and the Kazal peptides are similar to the vadodilator Vasotab (Takac et al., 2006; Xu et al., 2008; Zhang et al., 2014).

3.2.3. Small molecule binding domains—Tick and blood feeding insect sialomes abound with proteins that act as scavengers of hemostasis and inflammation agonists, named kratagonists (from the Greek kratos = to seize) (Ribeiro and Arca, 2009). In ticks and triatomines, the lipocalin family has been co-opted for this role, while in the Nematocera the D7 family, belonging to the odorant binding protein (OBP) superfamily (Valenzuela et al., 2002), has been co-opted. So far uniquely in sand flies, members of the Yellow protein family serve as serotonin kratagonists (Xu et al., 2011), and, uniquely in tabanids, members of the antigen-5 family are leukotriene seizers (Xu et al., 2012). The sialotranscriptome of *T. bromius* reveals members of the Yellow protein family, however with low expression (RRE=0.0013); this transcript could code for the canonical dopachrome conversion enzyme of housekeeping function (Johnson et al., 2001). Similarly, the deducted transcripts coding for lipocalins, juvenile hormone binding proteins, and members of the phosphatidylethanolamine binding proteins, all with RRE<0.07 may have housekeeping functions. However, Tb-3634 codes for an OBP having a Lys-Gly-Asp (KGD) motif flanked by cysteines, which could confer a disintegrin function and inhibit platelet aggregation (Suehiro et al., 1996). This protein is over 40% identical in aa sequence to mosquito and drosophilid homologs, all of which have the conserved Cys and KG, but not the aspartic acid residue.

3.2.4. Antigen 5 family—The CAP superfamily of proteins includes the Cysteine rich proteins of snake venoms, the Antigen-5 family of vespid venoms, and the Pathogenesis related proteins of plants (Gibbs et al., 2008). It is expressed in all sialomes analyzed to date. A salivary triatomine protein functions as a superoxide dismutase and inhibits platelet aggregation (Assumpcao et al., 2013). Uniquely in tabanids, *T. yao* tablysin has incorporated RGD and RTS domains flanked by disulfide bonds, inhibiting platelet aggregation and angiogenesis (An et al., 2011; Ma et al., 2010; Ma et al., 2011b; Xu et al., 2008). Surprisingly, the crystal structure of tablysin revealed a hydrophobic pocket, leading to the discovery that it acts as a leukotriene kratagonist (Xu et al., 2012). It is possible that this protein family has evolved in tabanids in a convergent manner to the triatomine lipocalins or mosquito D7 families, where individual members seize preferentially histamine, serotonin, norepinephrine or lipid agonists. The sialotranscriptome of *T. bromius* reveals several members of the tablysin family, several with large expression (RRE above 10 for 10 transcripts). Indeed, members of this family map 54% of the reads from the S class, indicating they account for over 50% of the mRNA coding for secreted products in the salivary glands of *T. bromius*.

Phylogenetic analysis of *T. bromius* and *T. yao* antigen-5 proteins and their closest matches from other Dipteran proteins (Supplemental fig 1) showed a robust clade of mosquito and

one *T. bromius* sequence that has a double antigen-5 domain (named “Double domain” in Supplemental fig 1), one robust clade containing *Drosophila*, black fly and mosquito sequences, named “Common” in Supplemental fig 1, and three clades of exclusively tabanid sequences. The two first named clades are associated with *T. bromius* contigs having low expression, while the tabanids exclusive clades have *T. bromius* transcripts of high expression. These three clades are named “Tab Fib” because some of the *T. yao* sequences are annotated as fibrinogenolytic enzymes, “Tab RGD-N” as most sequences have a N terminal RGD motif as detailed in previous tablysin studies (Ma et al., 2011b; Xu et al., 2012), and “Tab RGD-C” containing sequences with a RGD motif on the protein carboxyterminus. Within the RGD-C clade there is a *T. yao* protein with both RGD motifs. These different clades beg the questions of whether they share agonist binding pockets as found for tablysin, and whether the C terminus RGD motif could be functional. True disintegrins contain a RGD or KGD motif flanked by cysteines, the motif being at the edge of a hairpin loop that interacts with their integrin targets (Francischetti, 2010). The crystal structure of tablysin-15 showed that its RGD motif is located at the apex of a loop near the N-terminus of the protein (Supplemental fig 3A) (Xu et al., 2012). The loop is very similar in conformation to the homologous region in the wasp venom antigen Ves-5 and is stabilized by two disulfide bonds that are also conserved between the two proteins. Ves-5 does not contain an RGD motif, but the similarity in conformation and disulfide bonding features with tablysin-15 suggests that the conformation of this insertion loop served as a preadaptation for the evolution of an integrin antagonist function in CAP domain proteins of this type.

CAP domain proteins having the C-terminal RGD motif have not been tested for the presence of disintegrin activity. Molecular modeling of one of these (TABYAO241914335) using homology methods based on the structure of tablysin-15 was used to locate the structural position of this motif and to evaluate the likelihood of functionality. In the model, the C-terminal RGD sequence lies in the hinge region of the protein on the C-terminal side of β -strand 3 at the apex of a loop that could potentially allow it to interact with an integrin target (Supplemental fig. 3C). Conformational changes in this loop could increase the projection distance from the surface of the protein. The loop would likely be stabilized by a disulfide bond in the hinge region of the protein that links Cys-206 with Cys-217 (Supplemental fig. 3C). While the positioning of the RGD motif is compelling, evaluation of the activity of recombinant proteins would be required to verify this functional assignment.

The lipid binding pocket of tablysin-15 lies between α -helices 1 and 3 in the CAP domain and has been shown to accommodate a molecule of cysteinyl leukotriene, an eicosanoid modulator of inflammation and vascular tone (Supplemental fig. 3B). Alignment of residues occurring in the binding pocket shows that tablysin, and a number of its relatives, contain glycine at position 131 in the mature sequence (marked with a red arrow in Supplemental fig. 2) while the majority of sequences from *T. yao* and *T. bromius* contain tryptophan. Additionally, Ala-161 of tablysin is replaced by a bulkier residue, either leucine, threonine or methionine, in the majority of sequences. Examination of the model of the C-terminal RGD protein described above reveals that these bulky side chains fill most of the binding pocket, and that any members of the *Tabanus* antigen V family having these substitutions

would not have a lipid binding pocket lying between helices 1 and 3 (Supplemental fig. 3E). *T. bromius* sequence Tb-16383 matches tablysin at these positions and molecular modeling, using tablysin as a template, shows a well-formed binding pocket, strongly suggesting that this protein would also bind eicosanoids of the cysteinyl leukotriene type (Supplemental fig. 3D).

3.2.5. Immunity-related—This class of proteins is found in all sialomes, where they may help to control bacterial growth in the ingested food. TbSigP-212 has 66% identity to *Glossina morsitans morsitans* protein annotated as C-type lysozyme, and matches the LYZ1 - C-type lysozyme model of the CDD database with an e value of 4e-50.

Defensins are very divergent antimicrobial peptides having a conserved structure maintained by three disulphide bridges in arthropods (Tassanakajon et al., 2015). TbSigP-16379 is 88% identical to *T. yao* defensin TY 2. However, it is only 31% identical to *Lutzomyia longipalpis* defensin (gi|149728143). *Tabanus* defensins are shorter than other dipteran defensins, TbSigP-16379 mature peptide having 49 aa. Notice that most defensins shown in Supplemental figure 3 have the KR motif indicative of cleavage of the propeptide, which does not exist on the tabanids defensins, as well as in a few other peptides, indicating a different processing mechanism of the mature peptide. The peptides lacking the KR motif have a gap between the signal peptide cleavage site and the first conserved cysteine. The alignment of dipteran defensins (Supplemental figure 4) allows detection of the pattern C-x(4,12)-C-x(3)-C-x(7,8)-G-x-C-x(4,5)-C-x-C containing six conserved cysteines and one conserved glycine residue. When this pattern was used to search 986,619 sequences from Diptera organisms available in GenBank, 1,264 matches are found, 664 of which have 100 or less aa, indicating its robustness for searching defensins. TbSigP-16379 is well expressed, with a RRE=0.48, having been assembled from 29,356 reads.

Several lectins were identified, including 3 galectins and one C-type lectin. These have relatively low expression and may have a housekeeping function.

3.2.6. Tabanid-specific families—Tabimmunregulin peptides were identified in *T. yao* and *T. pleskei* and their action as modulators of lymphocyte function characterized (Xu et al., 2008; Zhao et al., 2009). Alignment of mature peptides from *T. bromius*, *T. yao*, *T. pleskei* and *Hybomitra bimaculata* (Supplemental figure 5) provides the pattern Q-x(6)-G-x(3)-K-G-x(3)-G-x(5)-P-x(4)-G. This pattern was used to search 986,619 sequences from Diptera organisms available in GenBank, retrieving only tabimmunregulins. As pointed out before (Xu et al., 2008), the RK duet indicative of cleavage is conserved in all but one sequence. Peptides isolated from fly salivary glands submitted to Edman degradation indicated the processing cleavage after RK motif (Xu et al., 2008; Zhao et al., 2009). As indicated above, *T. bromius* tabimmunregulins contain the most expressed contig in the transcriptome, TbSigP-17271 having accrued over 6 million reads.

A histidine-proline rich family of proteins with poor similarity to known proteins is also abundantly expressed, TbSigP-16493 having been assembled from over 300,000 reads, having an RRE=5.0. Histidine rich proteins of low complexity are found in ticks, mosquitoes, sand fly and black fly sialomes (Francischetti et al., 2009; Ribeiro et al., 2010),

and may function as antimicrobials possibly due to the repeated His residues that can chelate trace metals essential to microbial growth (De Smet and Contreras, 2005; Lai et al., 2004; Rydengard et al., 2006).

Three contigs code for related proteins annotated in Supplemental File 1 as mucin 30/88 family. These contigs are well expressed, TbSigP-17346 having and RER of 3.3. They have from 2 to 5 N-acetyl-galactosylation sites, and have poor similarity to known proteins.

Several contigs coding for short secreted peptides were identified. Most are poorly expressed, except for TbSigP-16616, coding for a mature peptide of 3.1 kDa assembled from 234,932 reads, and having a RER=3.87.

3.2.7. Conserved secreted proteins of unknown function—Supplemental File 1 displays several contigs coding for expanded families of conserved proteins of unknown function. Related proteins have been found before in salivary or midgut transcriptomes. It is possible that many of these proteins represent uncharacterized antimicrobial families. One of these families has related proteins in sand fly midgut and is quite well expressed, such as TbSigP-5660 (RER of 19.8), assembled from over 1 million reads.

3.2.8. Polymorphism analysis—A total of 1,426 CDS were found to have polymorphisms and a minimum base coverage of 30 (RPKM = 7) following read mapping as described in the methods section (Supplemental File S1). Transposable elements, the secreted and the unknown classes have the highest synonymous and non-synonymous rates (Table 4), as well as the highest rates of non-synonymous to synonymous mutations, further supporting the faster rate of evolution of salivary proteins from blood sucking arthropods (Arca et al., 2014).

4. Conclusions

The sialome of *Tabanus bromius*, together with the extensive previous analysis of *T. yao* (Ma et al., 2009; Xu et al., 2008) highlights the unique evolutionary pathways taken by tabanids to disarm their host's hemostasis and inflammation. Tabanids have uniquely evolved salivary tabimmunoregulins and other protein families of unknown function, and have uniquely modified the ubiquitous members of the antigen-5 family to confer them disintegrin and kratagonist functions. Further analysis of this family reveals members with additional disintegrin domains and agonist binding pockets indicative of functional differences acquired by gene expansion of the tablysin family. It should be remarked that these characteristics are different from those of other higher flies that evolved independently to hematophagy such as the tsetse *Glossina morsitans morsitans* (Alves-Silva et al., 2010), which do not sugar feed, have no salivary glycosidases and have large amounts on an inactive endonuclease with unknown function and various unique peptide families, or *Stomoxys calcitrans* (Wang et al., 2009) that have large amounts of a non-RGD antigen 5 family that binds immunoglobulins (Ameri et al., 2008) and have their own additional unique families. In common with other sialomes, apyrase, hyaluronidase, endonuclease, peroxidase, phospholipase, serine protease, dipeptidyl proteases, as well as protease inhibitors and immunity related transcripts are members of the salivary cocktail. Over three

thousand coding sequences were deposited to GenBank as a consequence of this work, adding to only two protein sequences currently available for *T. bromius* in Genbank. These CDS will help further work in the proteomic characterization of this species sialome, in future genomic annotation of this fly, and to develop specific immunological markers of host exposure to *T. bromius*.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, USA and by the Operational Program of Research and Development and co-financed with the European Fund for Regional Development (EFRD). Grant: ITMS 26240220030: Research and development of new biotherapeutic methods and its application in some illnesses treatment.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997; 25:3389–3402. [PubMed: 9254694]
- Alves-Silva J, Ribeiro JM, Van Den Abbeele J, Attardo G, Hao Z, Haines LR, Soares MB, Berriman M, Aksoy S, Lehane MJ. An insight into the sialome of *Glossina morsitans morsitans*. *BMC genomics.* 2010; 11:213. [PubMed: 20353571]
- Ameri M, Wang X, Wilkerson MJ, Kanost MR, Broce AB. An immunoglobulin binding protein (antigen 5) of the stable fly (Diptera: Muscidae) salivary gland stimulates bovine immune responses. *Journal of medical entomology.* 2008; 45:94–101. [PubMed: 18283948]
- An S, Ma D, Wei JF, Yang X, Yang HW, Yang H, Xu X, He S, Lai R. A novel allergen Tab y 1 with inhibitory activity of platelet aggregation from salivary glands of horseflies. *Allergy.* 2011; 66:1420–1427. [PubMed: 21848516]
- Arca B, Struchiner CJ, Pham VM, Sfera G, Lombardo F, Pombi M, Ribeiro JM. Positive selection drives accelerated evolution of mosquito salivary genes associated with blood-feeding. *Insect molecular biology.* 2014; 23:122–131. [PubMed: 24237399]
- Assumpcao TC, Ma D, Schwarz A, Reiter K, Santana JM, Andersen JF, Ribeiro JM, Nardone G, Yu LL, Francischetti IM. Salivary Antigen-5/CAP family members are Cu²⁺-dependent antioxidant enzymes that scavenge O₂⁻ and inhibit collagen-induced platelet aggregation and neutrophil oxidative burst. *The Journal of biological chemistry.* 2013; 288:14341–14361. [PubMed: 23564450]
- Birol I, Jackman SD, Nielsen CB, Qian JQ, Varhol R, Stazyk G, Morin RD, Zhao Y, Hirst M, Schein JE, Horsman DE, Connors JM, Gascoyne RD, Marra MA, Jones SJ. De novo transcriptome assembly with ABySS. *Bioinformatics (Oxford, England).* 2009; 25:2872–2877.
- Blasdell KR, Voysey R, Bulach D, Joubert DA, Tesh RB, Boyle DB, Walker PJ. Kotonkan and Obodhiang viruses: African ephemeroviruses with large and complex genomes. *Virology.* 2012; 425:143–153. [PubMed: 22305623]
- Blesa D, Martinez-Sebastian MJ. bilbo, a non-LTR retrotransposon of *Drosophila subobscura*: a clue to the evolution of LINE-like elements in *Drosophila*. *Molecular biology and evolution.* 1997; 14:1145–1153. [PubMed: 9364772]
- Bose R, Friedhoff KT, Olbrich S, Buscher G, Domeyer I. Transmission of *Trypanosoma theileri* to cattle by Tabanidae. *Parasitology research.* 1987; 73:421–424. [PubMed: 3658973]

- Chagas AC, Oliveira F, Debrabant A, Valenzuela JG, Ribeiro JM, Calvo E. Lundep, a sand fly salivary endonuclease increases *Leishmania* parasite survival in neutrophils and inhibits XIIa contact activation in human plasma. *PLoS Pathog.* 2014; 10:e1003923. [PubMed: 24516388]
- De Smet K, Contreras R. Human antimicrobial peptides: defensins, cathelicidins and histatins. *Biotechnology letters.* 2005; 27:1337–1347. [PubMed: 16215847]
- Dirie MF, Bornstein S, Wallbanks KR, Stiles JK, Molyneux DH. Zymogram and life-history studies on trypanosomes of the subgenus *Megatrypanum*. *Parasitology research.* 1990; 76:669–674. [PubMed: 2123546]
- Duckert P, Brunak S, Blom N. Prediction of proprotein convertase cleavage sites. *Protein Eng Des Sel.* 2004; 17:107–112. [PubMed: 14985543]
- Francischetti IM. Platelet aggregation inhibitors from hematophagous animals. *Toxicon.* 2010; 56:1130–1144. [PubMed: 20035779]
- Francischetti IMB, Sá-Nunes A, Mans BJ, Santos IM, Ribeiro JMC. The role of saliva in tick feeding. *Frontiers in Biosciences.* 2009; 14:2051–2088.
- Friend WG, Stoffolano JG. Feeding-behavior of the horsefly *Tabanus nigrovittatus* (Diptera, Tabanidae) - Effects of dissolved solids on ingestion and destination of sucrose or ATP diets. *Physiol Entomol.* 1991; 16:35–45.
- Galperin MY, Koonin EV. 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic acids research.* 2004; 32:5452–5463. [PubMed: 15479782]
- Gibbs GM, Roelants K, O'Bryan MK. The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins--roles in reproduction, cancer, and immune defense. *Endocrine reviews.* 2008; 29:865–897. [PubMed: 18824526]
- Hornok S, Micsutka A, Meli ML, Lutz H, Hofmann-Lehmann R. Molecular investigation of transplacental and vector-borne transmission of bovine haemoplasmas. *Veterinary microbiology.* 2011; 152:411–414. [PubMed: 21605950]
- Johnson JK, Li J, Christensen BM. Cloning and characterization of a dopachrome conversion enzyme from the yellow fever mosquito, *Aedes aegypti*. *Insect Biochem. Mol. Biol.* 2001; 31:1125–1135. [PubMed: 11520691]
- Julenius K, Molgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. *Glycobiology.* 2005; 15:153–164. [PubMed: 15385431]
- Karim S, Singh P, Ribeiro JM. A deep insight into the sialotranscriptome of the gulf coast tick, *Amblyomma maculatum*. *PLoS ONE.* 2011; 6:e28525. [PubMed: 22216098]
- Kazimirova M, Sulanova M, Kozanek M, Takac P, Labuda M, Nuttall PA. Identification of anticoagulant activities in salivary gland extracts of four horsefly species (Diptera, tabanidae). *Haemostasis.* 2001; 31:294–305. [PubMed: 11910198]
- Kazimirova M, Sulanova M, Trimmell AR, Kozanek M, Vidlicka L, Labuda M, Nuttall PA. Anticoagulant activities in salivary glands of tabanid flies. *Medical and veterinary entomology.* 2002; 16:301–309. [PubMed: 12243231]
- Kumar S, Tamura K, Nei M. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Briefings in bioinformatics.* 2004; 5:150–163. [PubMed: 15260895]
- Lai R, Takeuchi H, Lomas LO, Jonczy J, Rigden DJ, Rees HH, Turner PC. A new type of antimicrobial protein with multiple histidines from the hard tick, *Amblyomma hebraeum*. *Faseb J.* 2004; 18:1447–1449. [PubMed: 15247144]
- Lavoipierre MM. Feeding mechanism of blood-sucking arthropods. *Nature.* 1965; 208:302–303. [PubMed: 5882467]
- Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England).* 2010; 26:589–595.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing, S. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England).* 2009; 25:2078–2079.
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G,

- Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. 2012; 1:18. [PubMed: 23587118]
- Ma D, Gao L, An S, Song Y, Wu J, Xu X, Lai R. A horsefly saliva antigen 5-like protein containing RTS motif is an angiogenesis inhibitor. *Toxicon*. 2010; 55:45–51. [PubMed: 19635491]
- Ma D, Li Y, Dong J, An S, Wang Y, Liu C, Yang X, Yang H, Xu X, Lin D, Lai R. Purification and characterization of two new allergens from the salivary glands of the horsefly, *Tabanus yao*. *Allergy*. 2011a; 66:101–109. [PubMed: 20608917]
- Ma D, Wang Y, Yang H, Wu J, An S, Gao L, Xu X, Lai R. Anti-thrombosis repertoire of blood-feeding horsefly salivary glands. *Mol Cell Proteomics*. 2009; 8:2071–2079. [PubMed: 19531497]
- Ma D, Xu X, An S, Liu H, Yang X, Andersen JF, Wang Y, Tokumasu F, Ribeiro JM, Francischetti IM, Lai R. A novel family of RGD-containing disintegrins (Tablysin-15) from the salivary gland of the horsefly *Tabanus yao* targets $\alpha\text{IIb}\beta\text{3}$ or $\alpha\text{V}\beta\text{3}$ and inhibits platelet aggregation and angiogenesis. *Thrombosis and haemostasis*. 2011b; 105:1032–1045. [PubMed: 21475772]
- Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein engineering*. 1999; 12:3–9. [PubMed: 10065704]
- Reddy VB, Kouniga K, Mariano F, Lerner EA. Chrysoptin is a potent glycoprotein IIb/IIIa fibrinogen receptor antagonist present in salivary gland extracts of the deerfly. *J Biol. Chem*. 2000; 275:15861–15867. [PubMed: 10821849]
- Ribeiro JM. Blood-feeding arthropods: live syringes or invertebrate pharmacologists? *Infect Agents Dis*. 1995; 4:143–152. [PubMed: 8548192]
- Ribeiro JM, Assumpcao TC, Ma D, Alvarenga PH, Pham VM, Andersen JF, Francischetti IM, Macaluso KR. An insight into the sialotranscriptome of the cat flea, *Ctenocephalides felis*. *PLoS ONE*. 2012; 7:e44612. [PubMed: 23049752]
- Ribeiro JM, Chagas AC, Pham VM, Lounibos LP, Calvo E. An insight into the sialome of the frog biting fly, *Corethrella appendiculata*. *Insect biochemistry and molecular biology*. 2013
- Ribeiro JM, Mans BJ, Arca B. An insight into the sialome of blood-feeding Nematocera. *Insect biochemistry and molecular biology*. 2010; 40:767–784. [PubMed: 20728537]
- Ribeiro JM, Mather TN. *Ixodes scapularis*: salivary kininase activity is a metallo dipeptidyl carboxypeptidase. *Experimental parasitology*. 1998; 89:213–221. [PubMed: 9635445]
- Ribeiro JM, Valenzuela JG. Purification and cloning of the salivary peroxidase/catechol oxidase of the mosquito *Anopheles albimanus*. *J. Exp. Biol*. 1999; 202:809–816. [PubMed: 10069970]
- Ribeiro JMC, Arca B. From sialomes to the sialoverse: An insight into the salivary potion of blood feeding insects. *Adv Insect Physiol*. 2009; 37:59–118.
- Ribeiro JMC, Nussenzveig RH. The salivary catechol oxidase/peroxidase activities of the mosquito, *Anopheles albimanus*. *J. Exp. Biol*. 1993; 179:273–287. [PubMed: 8393473]
- Ribeiro JMC, Schwarz A, Francischetti IMB. A deep insight into the sialotranscriptome of the Chagas disease vector, *Panstrongylus megistus* (Hemiptera: Heteroptera). *J. Med. Entomol*. 2015; 52:351–358. [PubMed: 26334808]
- Rydengard V, Andersson Nordahl E, Schmidtchen A. Zinc potentiates the antibacterial effects of histidine-rich peptides against *Enterococcus faecalis*. *The FEBS journal*. 2006; 273:2399–2406. [PubMed: 16704414]
- Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic acids research*. 2003; 31:3381–3385. [PubMed: 12824332]
- Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. *Genome research*. 2009; 19:1117–1123. [PubMed: 19251739]
- Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol*. 1998; 6:175–182. [PubMed: 9783223]
- Suehiro K, Smith JW, Plow EF. The ligand recognition specificity of β3 integrins. *J Biol. Chem*. 1996; 271:10365–10371. [PubMed: 8626608]
- Takac P, Nunn MA, Meszaros J, Pechanova O, Vrbjar N, Vlasakova P, Kozanek M, Kazimirova M, Hart G, Nuttall PA, Labuda M. Vasotab, a vasoactive peptide from horse fly *Hybomitra bimaculata* (Diptera, Tabanidae) salivary glands. *J Exp Biol*. 2006; 209:343–352. [PubMed: 16391356]

- Tassanakajon A, Somboonwiwat K, Amparyup P. Sequence diversity and evolution of antimicrobial peptides in invertebrates. *Developmental and comparative immunology*. 2015; 48:324–341. [PubMed: 24950415]
- Taylor PD, Smith SM. Activities and physiological states of male and female *Tabanus sackeni*. *Medical and veterinary entomology*. 1989; 3:203–212. [PubMed: 2519664]
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 1997; 25:4876–4882. [PubMed: 9396791]
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols*. 2012; 7:562–578. [PubMed: 22383036]
- Valenzuela JG, Charlab R, Gonzalez EC, Miranda-Santos IKF, Marinotti O, Francischetti IM, Ribeiro JMC. The D7 family of salivary proteins in blood sucking Diptera. *Insect Mol. Biol*. 2002; 11:149–155. [PubMed: 11966880]
- Wang X, Ribeiro JM, Broce AB, Wilkerson MJ, Kanost MR. An insight into the transcriptome and proteome of the salivary gland of the stable fly, *Stomoxys calcitrans*. *Insect biochemistry and molecular biology*. 2009; 39:607–614. [PubMed: 19576987]
- Wilfert L, Jiggins FM. Flies on the move: an inherited virus mirrors *Drosophila melanogaster*'s elusive ecology and demography. *Molecular ecology*. 2014; 23:2093–2104. [PubMed: 24597631]
- Xu X, Francischetti IM, Lai R, Ribeiro JM, Andersen JF. Structure of protein having inhibitory disintegrin and leukotriene scavenging functions contained in single domain. *The Journal of biological chemistry*. 2012; 287:10967–10976. [PubMed: 22311975]
- Xu X, Oliveira F, Chang BW, Collin N, Gomes R, Teixeira C, Reynoso D, My Pham V, Elnaïem DE, Kamhawi S, Ribeiro JM, Valenzuela JG, Andersen JF. Structure and function of a “yellow” protein from saliva of the sand fly *Lutzomyia longipalpis* that confers protective immunity against *Leishmania major* infection. *The Journal of biological chemistry*. 2011; 286:32383–32393. [PubMed: 21795673]
- Xu X, Yang H, Ma D, Wu J, Wang Y, Song Y, Wang X, Lu Y, Yang J, Lai R. 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*. 2008; 7:582–590. [PubMed: 18087067]
- Yan X, Feng H, Yu H, Yang X, Liu J, Lai R. An immunoregulatory peptide from salivary glands of the horsefly, *Hybomitra atriperoides*. *Developmental and comparative immunology*. 2008; 32:1242–1247. [PubMed: 18502504]
- Zhang Z, Gao L, Shen C, Rong M, Yan X, Lai R. A potent anti-thrombosis peptide (vasotab TY) from horsefly salivary glands. *Int J Biochem Cell Biol*. 2014; 54:83–88. [PubMed: 25025626]
- Zhao QY, Wang Y, Kong YM, Luo D, Li X, Hao P. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. *BMC bioinformatics*. 2011; 12(Suppl 14):S2.
- Zhao R, Yu X, Yu H, Han W, Zhai L, Han J, Liu J. Immunoregulatory peptides from salivary glands of the horsefly, *Tabanus pleskei*. *Comparative biochemistry and physiology*. 2009; 154:1–5. [PubMed: 19332144]
- Zhu SY, Yi JP, Shen WD, Wang LQ, He HG, Wang Y, Li B, Wang WB. Genomic sequence, organization and characteristics of a new nucleopolyhedrovirus isolated from *Clanis bilineata* larva. *BMC genomics*. 2009; 10:91. [PubMed: 19243590]

Highlights

- The sialotranscriptome of the horse fly *Tabanus bromius* is disclosed.
- Over 75 million reads of 101 nt were assembled into 16,683 contigs.
- Over three thousand coding sequences were publicly deposited.
- Novel clades of antihemostatic proteins were discovered.
- Novel protein families are characterized.

Table 1

Functional classification of deduced coding sequences (CDS) from the sialotranscriptome of *Tabanus bromius*.

CDS Class	Total CDS	Reads	% CDS	% Reads	Reads / CDS
Secreted	320	37,610,426	7.8470	85.4902	117,533
Housekeeping	3,430	5,377,065	84.1099	12.2223	1,568
Unknown	218	942,753	5.3458	2.1429	4,325
Transposable elements	107	62,531	2.6238	0.1421	584
Viruses	3	1,067	0.0736	0.0024	356
Total	4,078	43,993,842	100	100	

Table 2

Functional classification of deduced coding sequences (CDS) coding for housekeeping proteins found in the sialotranscriptome of *Tabanus bromius*.

Class	CDS	Total reads	Reads / CDS	% Reads
Protein synthesis machinery	248	1,782,044	7,186	33.14
Protein modification	198	591,530	2,988	11.00
Unknown conserved	540	448,501	831	8.34
Signal transduction	422	330,922	784	6.15
Protein export	242	318,241	1,315	5.92
Energy metabolism	146	308,724	2,115	5.74
Transporters and channels	174	278,636	1,601	5.18
Transcription machinery	277	199,148	719	3.70
Unknown conserved membrane protein	165	189,246	1,147	3.52
Proteasome machinery	143	136,428	954	2.54
Cytoskeletal proteins	116	116,632	1,005	2.17
Nuclear regulation	138	105,189	762	1.96
Carbohydrate metabolism	112	96,482	861	1.79
Lipid metabolism	128	96,115	751	1.79
Transcription factor	19	78,860	998	1.47
Amino acid metabolism	48	68,475	1,427	1.27
Intermediary metabolism	36	44,530	1,237	0.83
Immunity	32	34,944	1,092	0.65
Detoxification	33	34,574	1,048	0.64
Nucleotide metabolism	59	32,732	555	0.61
Extracellular matrix	48	30,462	635	0.57
Oxidant metabolism/Detoxification	24	26,016	1,084	0.48
Storage	7	17,281	2,469	0.32
Nuclear export	12	10,724	894	0.20
Hormones	3	629	210	0.01
Total	3,430	5,377,065		100

Table 3

Functional classification of deducted coding sequences (CDS) coding for putative secreted proteins found in the sialotranscriptome of *Tabanus bromius*.

Class	CDS	Total reads	Reads / CDS	% Reads
Enzymes				
Glucosidases	5	594,462	118,892	1.58
Hyaluronidase	1	372,032	372,032	0.99
Peroxidase	1	280,371	280,371	0.75
Apyrase	3	506,525	168,842	1.35
Endonuclease	3	48,156	16,052	0.13
Phospholipase	8	205,203	25,650	0.55
Serine proteases	6	224,749	37,458	0.60
Dipeptidyl peptidase	4	118,056	29,514	0.31
Cathepsin L/B	7	7,411	1,059	0.02
Carboxypeptidase	3	2,116	705	0.01
Protease inhibitors				
Serpins	6	6,651	1,109	0.02
Kunitz family	10	15,385	1,539	0.04
Kazal/Vasotab	4	11,775	2,944	0.03
Cystatin	1	75	75	0.00
Cathepsin inhibitor	1	925	925	0.00
Metalloprotease inhibitor	1	559	559	0.00
Small molecule binding				
Antigen 5 family	25	20,402,883	816,115	54.25
Yellow protein family	1	84	84	0.00
Odorant/Pheromone binding protein family	5	2,407	481	0.01
Juvenile hormone binding protein	6	7,189	1,198	0.02
Lipocalin family	6	1,211	202	0.00
Phosphatidylethanolamine-binding protein	2	1,404	702	0.00
Immune related				
Defensin	1	29,356	29,356	0.08
Lysozyme	1	1,168	1,168	0.00
Lectins	4	4,048	1,012	0.01
Tabimmunregulin	3	10,959,162	3,653,054	29.14
Novel protein families				
HisPro rich protein family	6	665,783	110,964	1.77
Mucin 30/88	2	259,312	129,656	0.69
Other mucins	6	3,193	532	0.01
Short peptide family 30/2	18	2,423	135	0.01
Short peptide family 30/91	2	538	269	0.00
Short peptide family 30/95	3	1,617	539	0.00
Short peptide family 30/206	2	234,977	117,489	0.62

Class	CDS	Total reads	Reads / CDS	% Reads
Other short peptides	2	3,572	1,786	0.01
Conserved secreted proteins				
Family 30/19	5	9,801	1,960	0.03
Found in sand fly midgut	4	2,496,203	624,051	6.64
Other conserved secreted proteins	73	71,718	982	0.19
Other putative secreted proteins	79	57,926	733	0.15
Total	320	37,610,426		100

Table 4

Non synonymous (NS) and synonymous (S) substitution polymorphism in *Tabanus bromius* coding sequences according to their functional class. Only contigs with average read depth coverage > 30 were analyzed (FPKM > 7).

Class	Average NS × 100 / codons		SE	Average S × 100 / codons		SE	NS/S	N
Unknown	1.63	0.23		1.38	0.21		1.18	49
Secreted	1.29	0.16		1.33	0.14		0.96	86
Transposable element	1.30	0.21		2.63	0.41		0.49	27
Detoxification	0.79	0.16		1.72	0.30		0.46	27
Immunity	0.54	0.24		1.27	0.34		0.43	12
Cytoskeletal	0.34	0.08		1.06	0.20		0.32	36
Protein synthesis	0.46	0.07		1.44	0.16		0.32	79
Unknown conserved	0.41	0.05		1.45	0.08		0.28	257
Extracellular matrix	0.29	0.12		1.11	0.30		0.26	19
Protein modification	0.35	0.07		1.36	0.14		0.25	78
Nuclear regulation	0.34	0.09		1.52	0.22		0.22	51
Metabolism	0.32	0.03		1.47	0.09		0.22	217
Signal transduction	0.20	0.03		1.19	0.10		0.17	142
Transcription factor	0.26	0.13		1.59	0.30		0.16	26
Transcription machinery	0.21	0.04		1.32	0.12		0.16	99
Proteasome machinery	0.23	0.05		1.51	0.19		0.15	56
Protein export	0.18	0.04		1.32	0.11		0.14	98
Transporters and channels	0.18	0.03		1.33	0.16		0.14	59
Nuclear export	0.07	0.04		0.75	0.24		0.09	7

SE, standard error, N, number of coding sequences analyzed.