Ticks and Tick-borne Diseases xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Ticks and Tick-borne Diseases

journal homepage: www.elsevier.com/locate/ttbdis

An insight into the sialome of *Hyalomma excavatum*

José M.C. Ribeiro[a,*], Mirko Slovák[b], Ivo M.B. Francischetti[a]

[a] *Laboratory of Malaria and Vector Research, National Institute of Allergy and Infectious Diseases, Bethesda MD 20852, USA*
[b] *Institute of Zoology, Slovak Academy of Sciences, Dúbravská cesta 9, 84506, Bratislava, Slovakia*

ARTICLE INFO

Article history:
Received 25 April 2016
Received in revised form 23 August 2016
Accepted 24 August 2016
Available online xxx

Keywords:
Tick
Salivary gland
Saliva
Transcriptome
Sialome

ABSTRACT

Tick saliva contains hundreds or thousands of proteins that help blood feeding by impairing their hosts' hemostasis, inflammation and immunity. Salivary gland transcriptomes allow the disclosure of this pharmacologically active potion that consists of several multi-gene families, many of which are tick-specific. We here report the "de novo" assembly of ~138 million reads deriving from a cDNA library from salivary glands of adult male and female *Hyalomma excavatum* leading to the public deposition of 5337 coding sequences to GenBank. Among the deducted putative secreted proteins, metalloproteases, glycine rich proteins, mucins, anticoagulants of the madanin family and lipocalins were the most expressed. Novel protein families were identified. These sequences will permit proteomic studies aiming at identification of target antigens, epidemiological markers or salivary pharmaceuticals of interest, and contribute to our understanding of the fast evolution of the tick sialome.

Published by Elsevier GmbH.

1. Introduction

Tick saliva is represented by a complex assemblage of peptidic and non-peptidic compounds that assist blood feeding by disarming their host hemostasis and immunity (Chmelar et al., 2012; Francischetti et al., 2009; Kotal et al., 2015). Transcriptome studies indicate that ticks express hundreds or thousands of different polypeptides in their saliva that can be grouped into several multigene families, such as lipocalins, serpins, Kunitz-domain containing peptides and many other tick-specific families of unknown function (Chmelar et al., 2016; Francischetti et al., 2009). The diversity of these proteins is large, and probably derives from positive selection as their gene products benefit from mutations that evade, even temporarily, their hosts' immune system.

There are currently 64,150 protein sequences deposited on GenBank (as of Feb 11, 2016) annotated as deriving from tick salivary glands. While 26,370 are from the genus *Ixodes*, most from *I. ricinus*, the metastriate group counts 25,582 sequences for members of the *Amblyomma* genus and 11,513 for *Rhipicephalus*, most being for *R. pulchellus*. The genus *Hyalomma* is represented by only 100 sequences, 96 of which are from *H. marginatus rufipes*. In the present work we report the public deposition of 5337 sequences derived from the sialotranscriptome of adult male and female *H. excavatum* ticks that fed for different amounts of time on white rabbits. *H. excavatum* (Apanaskevich and Horak, 2005; Hoogstraal and Kaiser, 1959) is of veterinary importance as a vector of *Theileria* to cattle and sheep in Africa (Friedhoff, 1997), and is a suspected vector of Crimean-Congo hemorrhagic fever (Khan et al., 1997).

2. Material and methods

2.1. Ticks

Ticks were reared at the Institute of Zoology, Slovak Academy of Sciences. The original stock was the kind gift from Dr. Michael Samish, Kimron Veterinary Institute, Bait Dagan, Israel. White rabbits were used to rear all stages of this tick in the laboratory, as previously described (Slovak et al., 2002). The ticks were maintained in desiccators filled with saturated KCl solution to provide a RH of 85–90%, at a photoperiod of 16:8 (L:D) and temperature 24 + 2 °C. The salivary glands were dissected from adult ticks originated from F3 laboratory generation that were unfed or fed at different times as indicated in Table 1. Glands were stored in RNAlater (Qiagen, Valencia CA) until used for mRNA extraction. The glands were pooled before RNA extraction. The usage of animals in these experiments was approved by the State Veterinary and Food Administration of the Slovak Republic (permit numbers 928/10-221 and 1335/12-221).

2.2. RNA extraction, library preparation and sequencing

RNA preparation, library construction and sequencing were performed essentially as described previously (Ribeiro et al., 2014).

* Corresponding author.
E-mail addresses: jribeiro@nih.gov (J.M.C. Ribeiro), slovak.mirko@gmail.com (M. Slovák), ifrancischetti@niaid.nih.gov (I.M.B. Francischetti).

http://dx.doi.org/10.1016/j.ttbdis.2016.08.011
1877-959X/Published by Elsevier GmbH.

Table 1
Time of feeding, number of ticks and tick gender from which salivary glands were removed into RNAlater for library construction.

Fed	Females	Males
0	4	4
60–90 min	4	4
230–260 min	4	4
6-6.5 h	4	4
15 h	4	4
1 day	4	4
2 days	3	3
3 days	3	3
4 days	3	3
5 days	3	3
6 days	3	3
7 days	3	3
8 days	3	3
9 days – females dropped off	3	3
Total	48	48

mRNA library construction and sequencing were done by the NIH Intramural Sequencing Center. The salivary gland (SG) library was constructed using the TruSeq RNA sample prep kit, v2 (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 triatomine bug *Panstrongylus megistus*, the horse fly *Tabanus bromius* and the bat *Diphila ecaudata* were co-sequenced with *Hyalomma*, and we found that some cross-contamination of sequences occurred between the libraries (see below). Researchers reanalyzing the raw data should take this possibility of contamination into consideration. A total of 138,144,530 sequences of 101 nucleotides in length were obtained for the *Hyalomma* library. A paired-end protocol was used.

2.3. Bioinformatic tools used

The pipeline used has been described before (Ribeiro et al., 2015). Briefly, raw data were processed using RTA 1.12.4.2 and CASAVA 1.8.2. Reads were trimmed of low quality regions and 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 tenth from 21 to 91) and SOAPdenovo-Trans assembler (Luo et al., 2012). 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. Values of the reads per kilobase of transcript per million mapped reads (RPKM) (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 divided by the largest found number of reads mapped to a single CDS, which in the case of this transcriptome was a value of 1,354,561 mapped to a single madanin coding sequence. Automated annotation of proteins was based on a vocabulary of nearly 350 words found in matches to various databases, including Swissprot, Gene Ontology, KOG, Pfam, and SMART, Refseq-invertebrates and the acari subset of the GenBank sequences obtained by querying acari [organism] and retrieving all protein sequences. 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 aln (Li and Durbin, 2010) was used to map the reads to the CDS, producing SAI files that were joined by BWA sampe module, converted to BAM format, and sorted. The sequence alignment/map tools (samtools) package (Li et al., 2009) 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 20 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 (1000 iterations) of the phylogenies were done with the Mega package (Kumar et al., 2004).

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 PRJNA311286. This Transcriptome Shotgun Assembly project has been deposited at DDBJ/EMBL/GenBank under the accession GEFH00000000. The version described in this paper is the first version, GEFH01000000. Hyperlinked excel spreadsheets containing the coding sequences and their annotation are available at http://exon.niaid.nih.gov/transcriptome/Hexcav/Hyaexcav-web.xlsx (hyperlinked excel spreadsheet, 21 MB).

3. Results and discussion

3.1. Overview of the sialotranscriptome of Hyalomma excavatum

Following assembly of 138,144,530 reads, a total of 53,228 contigs were obtained (Supplemental file S1), from which we extracted 7875 coding sequences. These coding sequences mapped 57,440,028 reads, or 42% of the total reads. Their average length was 1125 nucleotides (nt) with 3273 CDS being equal or larger than 1000 nt. These CDS were classified into four classes: "secreted" (S), "housekeeping" (H), "unknown" (U) and "transposable elements" (TE) (Table 2). The S class had 1796 assigned CDS, and mapped 61% of all reads in accordance with the secretory nature of the organ. The H class produced 5511 CDS, mapping 36% of the reads. TE's accounted for 2.24% of the CDS and 0.94% of the reads, a typical finding when comparing to other sialotranscriptomes. Finally, 390 CDS were not able to be classified, representing 1.12% of the reads.

The housekeeping CDS were further classified by their function (Table 3), not surprisingly showing the category "protein synthesis" to be the most expressed and accruing 19% of the reads of the H