Pyrosequencing the salivary transcriptome of *Haemadipsa interrupta* (Annelida: Clitellata: Haemadipsidae): anticoagulant diversity and insight into the evolution of anticoagulation capabilities in leeches

Sebastian Kvist,1,a Mercer R. Brugler,2,3 Thary G. Goh,4 Gonzalo Giribet,1 and Mark E. Siddall2,3

1 Museum of Comparative Zoology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA
2 Division of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA
3 Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York 10024, USA
4 Institute of Biological Sciences, Science Faculty, University of Malaya, Kuala Lumpur 50603, Malaysia

Abstract. Freshwater, marine, and terrestrial leeches that depend on a diet of fresh blood have evolved salivary peptide components that inhibit normal thrombus formation by prey. Although bloodfeeding in leeches has long been of interest to biologists and medical practitioners alike, only a few studies have comprehensively examined the anticoagulant repertoires of hematophagous leeches, and these have largely been confined to representatives of Glossiphonidae, Hirudinidae, and Macrobdellidae. Here, we present 454 pyrosequencing data from the salivary transcriptome of the hematophagous terrestrial leech *Haemadipsa interrupta* (Haemadipsidae). Assembled transcripts were annotated using both similarity scores (BLAST) and gene ontology (Blast2GO). Subsequently, transcripts were examined within alignments containing well-characterized anticoagulants and other select bioactive (i.e., affecting the living cells of the prey) salivary proteins and phylogenies were reconstructed for each protein data set to verify orthology predictions. In total, transcripts significantly matching 20 salivary proteins of interest were found in the transcriptome, representing several different antagonistic pathways. After reviewing gene ontologies, alignments, and phylogenetic trees, sequences for 15 out of the 20 hits were deemed correctly annotated. Additionally, we recovered matches against several proteins that have previously been linked to anticoagulation (e.g., cathepsin and disintegrins), but the specific function of which in leeches needs further investigation. Finally, in light of these data as well as those previously published, we discuss our current understanding of the distribution and evolution of anticoagulants in leeches.

Additional key words: transcriptomics, Hirudinida, bioactive proteins, hematophagy

Thought to be a cure for a variety of ailments such as fever, hysteria, obesity, insomnia, ulcers, and an imbalance of the four humors (blood, phlegm, cholera, and melancholy), hematophagous leeches have been used in human phlebotomy for millennia (Sawyer 1981; Min et al. 2010; Elliott & Kutschera 2011). Concurrent with our growing understanding of the structure and function of leech anticoagulation factors that are responsible for persistent blood loss in patients, leeches have become more authoritatively used in modern medicine (Markwardt 1985; Adams 1988; Wells et al. 1993; Whitaker et al. 2004; Stange et al. 2012). Medicinal leeches are most frequently used as relievers of venous congestion to salvage microvascular free flaps or following digit replantation surgery. Their usefulness in this regard stems from their intense bloodfeeding behavior in combination with potent anticoagulation factors that they secrete into the wound (Salzet 2001). Although the European medicinal leech, *Hirudo verbana* CARENA 1820, has become the centerpiece for clinical hirudinology (Trontelj & Utevsky 2005; Siddall et al. 2007; Shain 2009), its
Anticoagulant repertoire has only recently been investigated in a comprehensive manner (Kvist et al. 2013); that study suggests that bioactive (i.e., affecting the living cells of the prey) salivary proteins with at least four different antagonistic pathways (factor Xa inhibitors, elastase inhibitors, plasmin inhibitors, and Kazal-type serine protease inhibitors) are expressed and secreted by the leech. In addition to H. verbana, salivary transcriptomes of other representatives of Hirudinidae and Macrobdeellidae have recently been screened for anticoagulation factors (Min et al. 2010; Siddall et al. 2011; Kvist et al. 2013), revealing an astonishing array of anticoagu-
lants even within a single individual. However, sev-
eral other leech families, terrestrial and aquatic, are hematophagous but largely or entirely unstudied in terms of the expressed anticoagulation factors. Specifically targeting the anticoagulant repertoires of terrestrial leeches will increase our understanding of the evolution of hirudinid bloodfeeding in general, and specific anticoagulation factors in particular. In this regard, the family Haemadipsidae is of special interest because of its biology and natural history.

Haemadipsidae includes terrestrial leeches, most of which are obligate bloodfeeders. The clade can be broadly divided into two main groups, two-jawed (duognathous) and three-jawed (trignathous) species. Recent phylogenetic reconstructions show that duognathous species represent a monophyletic group, nesting within trignathous species, rendering the latter paraphyletic (Borda & Siddall 2011). There are ~70 described haemadipsid species, which is likely an underestimate of their true diversity (Keegan et al. 1968; Richardson 1975; Sawyer 1986; Neymann & Sharma 1996, 2001; Borda & Siddall 2011). Roughly 24 of these are representatives of the trignathous genus Haemadipsa, whereas the remaining species have been placed in 30 (Nybelin 1943; Richardson 1975, 1978) or 15 different genera (Sawyer 1986; see Borda & Siddall 2011 for full discussion). Interestingly, the family shows an unusual biogeographic distribution, with the trignathous leeches found only in East and Southeast Asia and on the Indian subcontinent, and duognathous leeches found on Sahul, Wallacea, Melanesia, Oceania, Juan Fernandez, Madagascar, and the Seychelles. Haemadipsids are not known from Africa or South America; most other leech families have a global distribution. Haemadipsids are well known to field biologists of the Indo-Pacific tropics as small yet intensely aggressive leeches that infiltrate pants, boots, and shirts. In fact, no other group of annelids has inspired such passionate accounts by travelers or naturalists. They are typically found in “dank tropical jungles, misty ravines and showery, forested mountain-sides [where] they are among the most dominant and self-assertive elements” (Moore 1927). Because of their terrestrial preference, haemadipsid leeches have different dietary options than freshwater and marine leeches, and this provides a unique opportunity to compare how their mode of blood-feeding, as it pertains to anticoagulants, has evolved in general, and if it evolved in parallel with their choice of diet. That is, because of differing physiology and anatomy of their prey, it is likely that haemadipsid leeches possess chemical components in their pharmacological cocktails that differ from those of their aquatic counterparts.

Understanding the phylogenetic position of terres-
trial hematophagous leeches is particularly impor-
tant when investigating the evolution of bloodfeeding in Hirudinida. Using morphology, two nuclear ribosomal markers 18S rRNA and 28S rRNA, and mitochondrial 12S rRNA and cyto-
chrome c oxidase subunit I (COI), Borda & Siddall (2004) showed that Cycloboedellidae (also terrestrial leeches) is sister to the remaining hirudiniform leeches. Within the latter clade, Haemadipsidae is sister to the remaining families (Semiscolecidae, Xerobdellidae, Macrobdeellidae and Hirudinidae). In a later study with increased taxon sampling, Phillips & Siddall (2009) found a slightly shifted topology with Haemadipsidae grouping sister to Xerobdellidae and Hirudinidae, while Semiscolecidae (now including Macrobdeellidae) grouped as sister to that clade. As such, it appears as though terrestrial leeches occur toward the base of the evolutionary tree of jaw-bearing (arhynchobdellid) leeches. Inves-
tigations into the anticoagulant repertoires of hae-
madipsid leeches should be useful in elucidating both the presence of anticoagulants in early Hiru-
diniformes and potential specializations of this rep-
ertoire in response to the disparate prey diversity.

Methods

Specimen collection and cDNA synthesis

Leech specimens were collected from exposed skin during field studies in the tropical rainforest at the Ulu Gomback Field Study Centre located northeast of Kuala Lumpur, Malaysia (03°19′29.15″ N, 101°45′08.27″E), and were immediately preserved in RNAlater® (Qiagen, Valencia, CA, USA). Spec-
imens were identified as Haemadipsa interrupta (Moore 1935) using a stereomicroscope while submersed in RNAlater; identifications were later verified by molecular sequences. Prior to RNA
extraction, leeches were washed in 0.5% bleach for 1 min and subsequently rinsed in deionized water for 1 min to minimize contamination with surface bacteria. Using sterilized dissecting tools, salivary tissue masses (glandular tissue) from three leeches were removed aseptically and subsequently rinsed in 0.5% bleach for 1 min followed by rinsing in deionized water for 1 min. Total RNA then was isolated using a hybrid protocol based on that of Hale et al. (2009) and that of the RNeasy Plus Mini Kit (Qiagen, Valencia, CA, USA). A 1.5 mL microcentrifuge tube containing leech salivary gland tissue was submerged in liquid nitrogen, followed immediately by homogenization of the frozen tissue using a sterile pestle. To maintain the integrity of the RNA, 500 μL of TRIzol® (Life Technologies, Gaithersburg, MD, USA) was added as soon as the sample began to thaw; homogenization continued after the addition by transferring 22 mL of TRIzol. An additional 500 μL of TRIzol was then added to the 1.5 mL tube (1 mL in total), and tissue was further macerated using a sterile syringe. Immediately following maceration, the tube was vortexed and maintained at room temperature (RT) for 5 min. Thereafter, we added 200 μL of chloroform, inverted the tube for 20 s, and let it sit at RT for 3 min. The tube then was centrifuged at 12,800 g for 15 min at 4°C. The top colorless layer (containing RNA) was transferred to a new 1.5 mL tube; at this point, we transitioned from the protocol by Hale et al. (2009) to step 3 of the RNeasy® Plus Mini Kit (which begins by adding 600 μL of 70% ethanol to the tube, and subsequently transferring the sample to an RNeasy spin column) and followed the manufacturer’s protocols, including both optional steps (i.e., an extra spin to dry the membrane and repeat elution in RNase-free water [40 μL total]). The resulting concentration of total RNA was determined using an Agilent 2100 BioAnalyzer (Agilent Technologies, Palo Alto, CA, USA), and in particular, the Agilent RNA 6000 Nano Kit. First-strand cDNA synthesis was carried out in a 5 μL reaction consisting of 1.0 μ total RNA, 1 μL of 10 μmol L⁻¹ SMART IV oligo (5'-AAG CAG TGG TAT CAA CGC AGA GTG GCC ATT′cDNA synthesis primer; 5'-TAG AGG CCG AGG CCG CCG ACA TGG TGT TTT TTT TCT TTT TTT VN-3'), 1 μL of 10 μmol L⁻¹ MODIFIED CDS III (i.e., the 3' cDNA synthesis primer; 5'-AAG CAG TGG TAT CAA CGC AGA GTG GCC ATT′cDNA synthesis primer; 5'-TAG AGG CCG AGG CCG CCG ACA TGG TGT TTT TTT TCT TTT TTT VN-3'), and brought to final volume with RNase-free water. Contents were mixed, spun briefly, and incubated at 72°C for 2 min. Following incubation, the tube was immediately placed on ice for 2 min, spun briefly, and the following reagents were added to the tube: 2 μL of 5X First-Strand Buffer (Clontech Laboratories, Palo Alto, CA, USA), 1 μL of 20 mmol L⁻¹ DTT, 1 μL of 10 mmol L⁻¹ dNTP mix, and 1 μL of SuperScript II Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA) (10 μL total volume). Contents were mixed, spun briefly, and incubated at 42°C for 1.5 h. Thereafter, the tube was immediately placed on ice to terminate first-strand synthesis. The following reagents were combined to initiate second-strand cDNA synthesis: 2 μL of first-strand cDNA, 18.5 μL RNase-free water, 2.5 μL 10X Advantage 2 PCR Buffer, 0.5 μL of 10 mmol L⁻¹ dNTP mix, 0.5 μL of 10 μmol L⁻¹ 5' PCR Primer (5'-AAG CAG TGG TAT CAA CGC AGA GTG GCC ATT′cDNA synthesis primer; 5'-TAG AGG CCG AGG CCG CCG ACA TGG TGT TTT TTT TCT TTT TTT VN-3'), 0.5 μL of 10 mmol L⁻¹ CDS II/3' PCR Primer, and 0.5 μL 50X Advantage 2 Polymerase Mix (25 μL total volume). Contents were gently mixed, spun briefly, and placed in a preheated (95°C) thermocycler. The following thermal cycling program was used: 95°C for 1 min, 95°C for 15 s, 68°C for 6 min, repeat steps 2–3 for 21 cycles. Approximately 3 μL of second-strand cDNA was visualized on a 1.1% agarose gel. Second-strand cDNA then underwent SflI digestion by transferring 22 μL of second-strand cDNA to a 0.2 mL PCR tube, adding 2.79 μL SflI Buffer, 2.79 μL SflI enzyme, and 0.28 μL of 10 mg mL⁻¹ (=100X) BSA, and incubating at 50°C for 2 h. Lastly, the sample was run through a Qia-gen QIAquick PCR Purification Kit column (and eluted in Buffer EB), quantified on a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, USA), and visualized using an Agilent 2100 BioAnalyzer High Sensitivity DNA Kit.

Library construction, emPCR, and pyrosequencing

The cDNA Rapid Library (RL) was prepared using a Roche 454 GS RL Prep Kit by following manufacturer’s protocols for individual sample cleanup as outlined in the Roche 454 RL Preparation Method Manual (Roche Applied Sciences, Indianapolis, IN, USA). Using a Quantifluor™-ST Fluorometer (Promega, Madison, WI, USA) in combination with the RL Quantitation Calculator, we calculated a sample concentration of 4.21×10⁸ molecules μL⁻¹ (R²=0.99356), which was diluted to a working stock of 1×10⁷ molecules μL⁻¹. Emulsion-based clonal amplification (PCR) was carried out using the GS Junior Titanium emPCR (Lib-L) Kit and following manufacturer’s protocols as outlined in the emPCR Amplification Method Manual (Lib-L). This manual was also used for bead recovery, DNA library bead enrichment, and sequence primer annealing. Enriched beads were
prepared for sequencing on a GS Junior Titanium PicoTitrePlate Device using the GS Junior Titanium Sequencing Kit and following manufacturer’s protocols as outlined in the Sequencing Method Manual. Massively parallel single-end pyrosequencing was conducted on a 454 GS Junior at the Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY, USA.

Filtering, trimming, and assembly

Post-sequencing processing involved a multi-tiered approach to assure the quality of downstream sequence data, and was initiated using the five standard 454 quality filters on the GS Junior (Dot, Mixed, Signal Intensity, Primer and TrimBack Valley). Thereafter, sff_extract (http://bioinf.comav.upv.es/sff_extract/index.html) was used to create .fasta, .fasta.qual, .fastq, and .xml files. In addition, sff_extract trimmed key/adaptor sequences and removed low quality reads (i.e., any base listed in lower case). After visualizing the results of sff_extract using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), two binaries, FASTX_trimmer (-f 20 –l 500 –Q 33) and FASTQ_quality_trimmer (-v –l 25) (both part of the FASTX toolkit [http://hannonlab.cshl.edu/fastx_toolkit/]), were used to further trim low quality regions; only bases between positions 20–500 and those with a Phred quality score ≥25 were retained in the data set (–Q 33 indicates that the Phred quality values are standard 454/ Sanger scores [0–93] encoded as ASCI characters 33–126). After utilizing FASTX_trimmer and FASTQ_quality_trimmer, FastQC was again used to visualize and verify the overall quality of the reads.

As of July 2013, there were no available transcriptomes for species closely related to H. interrupta. Thus, a de novo assembly of the filtered reads was conducted using MIRA v. 3.4.0 (Chevreux et al. 1999); MIRA was called as follows: mira –project= MyFile –job=denovo,est,accurate,454 454_SETTINGS –CL:qc=no;cpat=no >&log_assembly.txt (where qc= quality clip; cpat=poly-A clipping).

Open reading frames and BLAST queries

To minimize the risk of introns embedded within the data, the program getorf, which is part of the EMBOSS package (Rice et al. 2000), was used to extract open reading frames (ORFs) greater than 200 nucleotides in length within the contigs. Getorf employed the standard genetic code, and only ORFs beginning with a methionine (Met) and ending in a stop codon were extracted.

Uninterrupted contigs then were annotated utilizing a series of BLAST searches against various databases (all BLAST searches were carried out using BLAST 2.2.22). First, all contigs were queried against three databases constructed from GenBank non-redundant protein sequences (nr), nucleotide sequences (nt), and expressed sequence tags (EST) using blastall (-p BLASTx against nr and BLASTn against nt and EST). Results of these searches then were used as authoritative cross-controls (reciprocal best hit approach: Ge et al. 2005; Fang et al. 2010) for BLASTx searches against a locally compiled data set of amino acids from both well-characterized anticoagulants (predominantly from leeches) and other select salivary proteins from previous leech transcriptome studies (Table 1). That is, the contigs that best matched the sequences in the local data set were also BLASTed against GenBank nr, nt, and EST to confirm that they did not match other proteins at a better e-value. Positive reciprocal hits involved significant matches of the contigs to sequences annotated as the same bioactive protein as its best match in the locally compiled data set. By contrast, negative reciprocal hits occurred when the contigs matched a different bioactive protein than the best hit against the local data set. In addition, the raw FilterPass reads were both queried against GenBank nt (via BLASTn) and against the assembled contigs using BLASTn to get a better understanding of the distribution and coverage of the reads and the expression levels of the proteins.

Amino acid sequences for each of the best matches (those that scored the lowest e-values) against the known salivary proteins then were aligned in a pairwise fashion with those of the respective archetypal sequence using MAFFT (Katoh et al. 2005) on the European Bioinformatics Institute website (http://www.ebi.ac.uk/Tools/msa/mafft/) applying default settings. The term archetypal here refers either to the corresponding well-characterized anticoagulant (but not always that from the first investigation of the respective protein in leeches) or other bioactive proteins that have previously been considered of interest based on their matches against salivary peptides (see, e.g., Min et al. 2010). Our rationale behind pairwise alignments, as opposed to multiple sequence alignments, is an in-depth comparison to the known and, in most cases, functional anticoagulant sequence. Whereas multiple sequence alignments could be used to understand mutational rates across clades or selection pressures acting on the proteins, here we merely attempt to annotate sequences based on their similarity and composition as compared to a single,
Table 1. Anticoagulants included in the locally compiled data set used for the BLASTx comparisons.

<table>
<thead>
<tr>
<th>Organism</th>
<th>Bioactive protein</th>
<th>Antagonistic pathway</th>
<th>Genbank accession number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Hirudin</td>
<td>Thrombin inhibitor</td>
<td>Q07558</td>
<td>Scacheri et al. (1993)</td>
</tr>
<tr>
<td><em>Poecilobdella viridis</em></td>
<td>Hirudin</td>
<td>Thrombin inhibitor</td>
<td>P84590</td>
<td>Vankhede et al. (2005)</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Hirudin II</td>
<td>Thrombin inhibitor</td>
<td>P28504</td>
<td>Scharf et al. (1989)</td>
</tr>
<tr>
<td><em>Hirudinaria manillensis</em></td>
<td>Hirullin</td>
<td>Thrombin inhibitor</td>
<td>P26631</td>
<td>Tulinsky &amp; Qiu (1993)</td>
</tr>
<tr>
<td><em>Haemadipsa sylvestris</em></td>
<td>Unidentified</td>
<td>Thrombin inhibitor</td>
<td>CAA79672</td>
<td>Strube et al. (1993)</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Bdellin</td>
<td>Trypsin, plasmin inhibitor</td>
<td>P09865</td>
<td>Fink et al. (1986)</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Destabillase I</td>
<td>Destabilized fibrin de-polymerase</td>
<td>AAA96144</td>
<td>Zavalova et al. (1996)</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Destabillase II</td>
<td>Destabilized fibrin de-polymerase</td>
<td>AAA96143</td>
<td>Zavalova et al. (1996)</td>
</tr>
<tr>
<td><em>Theromyzon tessulatum</em></td>
<td>Cystatin</td>
<td>Cystein proteinase inhibitor</td>
<td>AAN28679</td>
<td>Lefebvre et al. (2004)</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Eglin c</td>
<td>Leukocyte elastase, cathepsin G inhibitor</td>
<td>0905140A</td>
<td>Knecht et al. (1983)</td>
</tr>
<tr>
<td><em>Haementeria officinalis</em></td>
<td>LAPP</td>
<td>Platelet aggregation inhibitor</td>
<td>Q01747</td>
<td>Keller et al. (1992)</td>
</tr>
<tr>
<td><em>Macrobdella decora</em></td>
<td>Decorsin</td>
<td>Glycoprotein IIb/IIIa inhibitor</td>
<td>P17350</td>
<td>Seymour et al. (1990)</td>
</tr>
<tr>
<td><em>Theromyzon tessulatum</em></td>
<td>Therostasin</td>
<td>Factor Xa inhibitor</td>
<td>Q9NBW4</td>
<td>Chopin et al. (2000)</td>
</tr>
<tr>
<td><em>Haementeria ghilianii</em></td>
<td>Ghilanten</td>
<td>Factor Xa inhibitor</td>
<td>AAB21233</td>
<td>Brankamp et al. (1991)</td>
</tr>
<tr>
<td><em>Haementeria ghilianii</em></td>
<td>Ghilanten</td>
<td>Factor Xa inhibitor</td>
<td>P16242</td>
<td>Blankenship et al. (1990)</td>
</tr>
<tr>
<td><em>Haementeria officinalis</em></td>
<td>Antistasin</td>
<td>Factor Xa inhibitor</td>
<td>AAA29193</td>
<td>Han et al. (1989)</td>
</tr>
<tr>
<td><em>Haementeria officinalis</em></td>
<td>Antistasin</td>
<td>Factor Xa inhibitor</td>
<td>P15358</td>
<td>Han et al. (1989)</td>
</tr>
<tr>
<td><em>Hirudo nipponia</em></td>
<td>Guamerin</td>
<td>Factor Xa (elastase) inhibitor</td>
<td>AAD09442</td>
<td>Jung et al. (1995)</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Hirustasin</td>
<td>Thrombin, trypsin, chymotrypsin, cathepsin G, kallikrein inhibitor</td>
<td>P80302</td>
<td>Söllner et al. (1994)</td>
</tr>
<tr>
<td><em>Haementeria officinalis</em></td>
<td>Saratin</td>
<td>Platelet aggregation inhibitor</td>
<td>2K13-X</td>
<td>Gronwald et al. (2008)</td>
</tr>
<tr>
<td><em>Macrobdella decora</em></td>
<td>Ficolin</td>
<td>Complementary cascade stimulator, immunoresponse</td>
<td>N/A</td>
<td>Min et al. (2010); cluster 686</td>
</tr>
<tr>
<td><em>Hirudo medicinalis</em></td>
<td>Leech derived tryptase inhibitor</td>
<td>Tryptase, trypsin, chymotrypsin inhibitor</td>
<td>AAB33769</td>
<td>Sommerhoff et al. (1994)</td>
</tr>
<tr>
<td><em>Placobdella ornata</em></td>
<td>Ornatin</td>
<td>Glycoprotein IIb/IIIa inhibitor</td>
<td>AAB28218</td>
<td>Mazur et al. (1993)</td>
</tr>
<tr>
<td><em>Haemadipsa sylvestris</em></td>
<td>Haemadin</td>
<td>Thrombin inhibitor</td>
<td>AAA15044</td>
<td>Strube et al. (1993)</td>
</tr>
<tr>
<td><em>Hirudinaria manillensis</em></td>
<td>Bufrudin</td>
<td>Thrombin inhibitor</td>
<td>P81492</td>
<td>Scacheri et al. (1993)</td>
</tr>
<tr>
<td><em>Theromyzon tessulatum</em></td>
<td>Theromin</td>
<td>Thrombin inhibitor</td>
<td>P82354</td>
<td>Salzet et al. (2000)</td>
</tr>
<tr>
<td><em>Haementeria ghilianii</em></td>
<td>Haementin</td>
<td>Fibrinogenolytic protease</td>
<td>Q7M3P9</td>
<td>Swadesh et al. (1990)</td>
</tr>
<tr>
<td><em>Hirudinaria manillensis</em></td>
<td>Gelin</td>
<td>Elastase, cathepsin G, chymotrypsin, subtilisin inhibitor</td>
<td>AAB27871</td>
<td>Electricwala et al. (1992)</td>
</tr>
<tr>
<td><em>Hirudo nipponia</em></td>
<td>Piguamerin</td>
<td>Factor Xa, trypsin, kallikrein inhibitor</td>
<td>P81499</td>
<td>Kim &amp; Kang (1998)</td>
</tr>
</tbody>
</table>
established sequence. The alignments then were visualized using Jalview ver. 2 (Waterhouse et al. 2009), and prediction of secretory signal peptides was performed on the SignalP 4.1 (Petersen et al. 2011) server at http://www.cbs.dtu.dk/services/SignalP/. Following alignment, in cases where the sequences of the archetypal bioactive proteins were substantially longer than the contigs, these were truncated at the 5′ end, 3′ end, or both. Moreover, in some cases, the contigs were truncated in the 3′-end. This was done to investigate the conservation and divergences in only the sequence regions that were represented by orthologous regions in both the archetypal protein and the contigs.

Blast2GO (Conesa et al. 2005) was used to infer functional annotations using gene ontology. The assembled transcripts were jointly submitted to Blast2GO to clarify the proportional allocation of the partial transcriptome to various classification categories. The resulting GO terms were subsequently interpreted using the CateGOrizer (Hu et al. 2008) platform at http://www.animalgenome.org/bioinfo/tools/countgo/, and pie charts illustrating the proportions were created in Microsoft Excel.

**Phylogenetic analyses**

Putative orthology and evolutionary relationships within each cluster of select bioactive salivary proteins were also examined by creating phylogenetic hypotheses of aligned amino acid sequences. In addition to the sequences of the best scoring contig and the archetypal bioactive protein, the final matrices included amino acid sequences employed by Kvist et al. (2013) in a similar study on the evolution of bloodfeeding in three species of medicinal leeches. Amino acid sequences were aligned using MAFFT employing the G-INS-i strategy with a gap-opening cost of 3.0 and default settings for all other parameters. Phylogenetic analyses were conducted in TNT (Goloboff et al. 2008) under the parsimony criterion, where a traditional search was performed for each data set employing 100 initial addition sequences and TBR branch swapping. All characters were unweighted and non-additive and gaps were treated as missing data. Because of the lack of appropriate outgroups, all trees were left unrooted.

For the leech derived tryptase inhibitor (LDTI) data set, which was not represented in the study by Kvist et al. (2013), a potentially orthologous protein sequence for aspolin (accession number AB189965; see below) was included and the following accessions were downloaded from GenBank and added to the matrix: P80424, 1LDT_L, 1LDT_T, 2KMR_A, 2KMQ_A, 2KMO_A, 2KMP_A. The best scoring contig matching bufrudin (also not represented in the study by Kvist et al. 2013), as well as the archetypal sequence (accession number P81492), was added to the hirudin data set. Unfortunately, too few sequences are available on GenBank for leech-derived cystatin and theromin to constitute workable data sets for phylogenetic inferences (only a single sequence of each). Therefore, phylogenies were not estimated for these data sets. Similarly, phylogenies were not estimated for the His-rich proteins (as their affiliation in terms of protein family and their function remains unknown) or the Kazal-type serine-protease inhibitors.

**Results**

**Raw reads: BLAST, trimming, and assembly**

Pyrosequencing of the cDNA library resulted in 200,316 Raw wells, 190,511 KeyPass wells, and 141,765 (74.4%) FilterPass reads (the latter of which passed all five quality filters). Average read length was 427.5±109.8 base pairs with a total of
BLAST and Blast2GO annotations

Complete lists of all results from the BLASTx and BLASTn searches using the contigs as queries are presented in Tables S1 and S2, respectively. On the one hand, the BLASTn searches using the raw reads as queries indicated that fully 16,062 reads found no match in the nt database, 2,100 matched leech nuclear 18S rDNA at 1e^-10 or below, 746 matched leech nuclear 28S rDNA at 1e^-10 or below, and 2,843 matched leech mtDNA transcripts including 12S rDNA, 16S rDNA, ATP6, Cox1, Cox2, Cox3, Cytb, ND1, ND4, and ND6. On the other hand, only a very small number of the assembled contigs scored significantly against sequences annotated as mitochondrial, suggesting that MIRA efficiently trimmed these data from the data set during assembly. Our initial morphological identification of the specimens was confirmed by several contig BLASTn and BLASTx hits matching sequences derived from Haemadipsa interrupta (e.g., c232 matching 18S rDNA from H. interrupta at 1e^-127).

Through BLASTx searches against the locally compiled data set, unique ORFs matching the following proteins at an e-value of 1e^-5 or below were found in the H. interrupta transcriptome (Table 2): cystatin, manillase, an unidentified thrombin inhibitor, leukocyte elastase inhibitor, lectin-like c-type lectin, antistasin, ficolin, guamerin, piguamerin, bdellin, Kazal-type serine protease inhibitors, gilantin, HIS-rich proteins, therostasin, leech-derived trypsin inhibitor (LDTI), bufrudin, eglin c, saratin, theromin, and leech antiplatelet protein (LAPP). After reciprocally BLASTing sequences against GenBank nr, nt, and EST databases, nine of the best 20 bioactive protein-matching ORFs show identical annotations to those of the best reciprocal BLAST(x and/or n) hit; for six of the ORFs, the best reciprocal BLAST(x and/or n) hits were against unannotated sequences; for four of the ORFs, the best reciprocal BLAST(x and/or n) hits were against sequences annotated for the same family of proteins as the best BLASTx hit against the local data set; and for only one ORF did the best reciprocal BLASTx and BLASTn hit match a sequence annotated as a different anticoagulant or other bioactive protein but at a much higher e-value (Table 2). Using the nr database, no significant reciprocal BLASTx hit was found for contig c3900_1, which initially matched a highly expressed HIS-rich protein derived from the salivary glands of the North American medicinal leech Macrobdella decora (SAY 1824) (see Table 2). In addition, the BLASTn searches against GenBank EST offered no further information concerning annotations as all contigs, save for c5389_2, matched unannotated sequences; also note that most of the hits were against hirudinid taxa. Several e-values resulting from the reciprocal BLASTx/n searches were insignificant, in that they were several orders of magnitude higher than the best match found against the locally compiled database of bioactive proteins, regardless of the annotation. In addition to the matches against well-characterized anticoagulants and other bioactive salivary proteins, numerous ORFs returned significant hits against proteins with putative involvement in anticoagulation activity, such as cathepsin, disintegrins, galactose-binding lectin, and leech carboxypeptidase inhibitors. When mapping the raw reads back to the putatively anticoagulation-related contigs, more than half of the reads that found a significant match mapped against disintegrins (185 raw reads), antistasins (167 raw reads), and c-type/galactose-binding lectin (144 raw reads), and this may serve as a coarse, initial indication of expression levels within the transcriptome. The number of raw reads mapped against each of the bioactive proteins is listed in Table 2.

Blast2GO and CateGOrizer mapped the contigs to 75 of the MGI_GO_slim1 ancestor terms by single count; 17 contigs found no match in the GO category database. Other biological process, other molecular function, and other cellular component were the most frequently encountered hits for each of the three categories (biological process, molecular function, and cellular component, respectively), indicating that the gene ontology for several of our proteins remains largely unknown (Fig. 1). The structures and functions of these BLAST and Blast2GO annotated, putative anticoagulation-related proteins, are discussed further below.
<table>
<thead>
<tr>
<th>Contig</th>
<th>Length (nt)</th>
<th>Best anticoagulant BLASTx hit (e-value)</th>
<th>% identity(^a)</th>
<th>Signal peptide</th>
<th>Raw reads mapped back</th>
<th>Best reciprocal BLASTx hit against NCBI nr (e-value)</th>
<th>Best reciprocal BLASTn hit against NCBI nt (e-value)</th>
<th>Best reciprocal BLASTn hit against NCBI EST (e-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c3896_2</td>
<td>303</td>
<td>Cystatin (2e-32)</td>
<td>54.16</td>
<td>No</td>
<td>47</td>
<td>AAN28679 Cystatin B (8e-34)</td>
<td>EU583802 cystatin B-like protein (5e-07)</td>
<td>FP616309 Hirudo medicinalis unannotated sequence (6e-43)</td>
</tr>
<tr>
<td>c1130_2</td>
<td>465</td>
<td>Manillase (1e-30)</td>
<td>48.91</td>
<td>No</td>
<td>52</td>
<td>ELT88779 unannotated sequence (7e-06)(^b)</td>
<td>XM003441864 zinc finger protein 281-like (0.059)(^d)</td>
<td>JZ184258 Asiaticobdella fenestrata unannotated sequence (7e-51)</td>
</tr>
<tr>
<td>c7044_1</td>
<td>258</td>
<td>Thrombininhibitor (8e-29)</td>
<td>75.81</td>
<td>Yes</td>
<td>66</td>
<td>CAA79672 thrombininhibitor (3e-25)</td>
<td>Z19864 thrombininhibitor (2e-37)</td>
<td>FL388489 Zea mays unannotated sequence (0.095)</td>
</tr>
<tr>
<td>c729_5</td>
<td>693</td>
<td>Leukocyte elastase inhibitor (1e-27)</td>
<td>31.77</td>
<td>No</td>
<td>24</td>
<td>XP003225626 leukocyte elastase inhibitor-like (2e-41)</td>
<td>NM026460 serine (or cysteine) peptidase inhibitor (3e-08)</td>
<td>FP611221 Hirudo medicinalis unannotated sequence (4e-43)</td>
</tr>
<tr>
<td>c14817_4</td>
<td>498</td>
<td>C-type lectin (8e-27)</td>
<td>37.50</td>
<td>Yes</td>
<td>144</td>
<td>XP002588290 unannotated sequence (8e-13)(^c)</td>
<td>XM003760904 Fc fragment of IgE (1e-04)</td>
<td>FP603066 Hirudo medicinalis unannotated sequence (4e-04)</td>
</tr>
<tr>
<td>c5961_1</td>
<td>471</td>
<td>Antistasin (4e-25)</td>
<td>40.15</td>
<td>Yes</td>
<td>167</td>
<td>AAA29193 antistasin (4e-15)</td>
<td>XM001528584 conserved hypothetical protein (0.21)</td>
<td>JZ185125 Asiaticobdella fenestrata unannotated sequence (3e-17)</td>
</tr>
<tr>
<td>c1477_1</td>
<td>270</td>
<td>Ficolin (1e-21)</td>
<td>66.07</td>
<td>No</td>
<td>6</td>
<td>CBM41043 fibrinogen-related protein 4 (2e-28)</td>
<td>AL60365 unannotated sequence (8e-10)</td>
<td>FP612077 Hirudo medicinalis unannotated sequence (5e-31)</td>
</tr>
<tr>
<td>c14138_1</td>
<td>225</td>
<td>Guamerin (8e-17)</td>
<td>62.50</td>
<td>Yes</td>
<td>49</td>
<td>P82107 Bdellastasin (1e-20)</td>
<td>U38282 guamerin (2e-03)</td>
<td>JK310652 Enchytraeus albidas unannotated sequence (5e-04)</td>
</tr>
<tr>
<td>c6218_1</td>
<td>252</td>
<td>Piguamerin (5e-16)</td>
<td>57.78</td>
<td>Yes</td>
<td>55</td>
<td>P81499 Piguamerin (4e-08)</td>
<td>AL627327 unannotated sequence (0.36)(^d)</td>
<td>FN136614 Homo sapiens unannotated sequence (0.093)</td>
</tr>
<tr>
<td>c2073_1</td>
<td>207</td>
<td>Bdellin (2e-14)</td>
<td>64.29</td>
<td>Yes</td>
<td>28</td>
<td>AAF73890 Bdellin-KL (9e-11)</td>
<td>FR718673 unannotated sequence (1.00)</td>
<td>FP666172 Hirudo medicinalis unannotated sequence (5e-04)</td>
</tr>
<tr>
<td>c4204_4</td>
<td>414</td>
<td>Kazal-type serpin (2e-13)</td>
<td>39.29</td>
<td>No</td>
<td>15</td>
<td>EKC29034 A disintegrin and metalloproteinase with thrombospondin motifs (0.37)</td>
<td>AL662954 unannotated sequence (0.052)</td>
<td>EC743006 Polytomella parva unannotated sequence (0.004)</td>
</tr>
</tbody>
</table>

\(^a\) Percentage identity with the BLASTx hit.

\(^b\) Cystatin.

\(^c\) Thrombininhibitor.

\(^d\) Leukocyte elastase inhibitor-like.

\(^e\) C-type lectin.

\(^f\) Antistasin.

\(^g\) Ficolin.

\(^h\) Guamerin.

\(^i\) Piguamerin.

\(^j\) Bdellin.

\(^k\) Kazal-type serpin.

\(^l\) Cystatin.

\(^m\) Thrombininhibitor.

\(^n\) Leukocyte elastase inhibitor-like.

\(^o\) C-type lectin.

\(^p\) Antistasin.

\(^q\) Ficolin.

\(^r\) Guamerin.

\(^s\) Piguamerin.

\(^t\) Bdellin.

\(^u\) Kazal-type serpin.
<table>
<thead>
<tr>
<th>Contig</th>
<th>Length (nt)</th>
<th>Best anticoagulant BLASTx hit (e-value)</th>
<th>Best reciprocal BLASTx hit against NCBI nr (e-value)</th>
<th>Best reciprocal BLASTn hit against NCBI nt (e-value)</th>
<th>Best reciprocal BLASTn hit against NCBI EST (e-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>c3999_2</td>
<td>459</td>
<td>LDTI (4e-13)</td>
<td>AAK58688 non-classical Kazal type inhibitor bdellin-KL (4e-10)</td>
<td>AB189965 aspartic-rich protein aspolinin (9e-19)</td>
<td>ES379288 Poecilia reticulata unannotated sequence (2e-19)</td>
</tr>
<tr>
<td>c3664_1</td>
<td>264</td>
<td>Ghilanten (4e-10)</td>
<td>XP004082650 cysteine-rich motor neuron 1 protein-like (1e-04)</td>
<td>NG011451 integrin, alpha X (0.11)</td>
<td>AA372712 Homo sapiens unannotated sequence (0.097)</td>
</tr>
<tr>
<td>c3900_1</td>
<td>258</td>
<td>HIS-rich protein (2e-08)</td>
<td>No significant similarity found</td>
<td>XM001233735 unannotated sequence (0.11)</td>
<td>GW218473 Coccomyxa sp. unannotated sequence (4.1)</td>
</tr>
<tr>
<td>c5389_2</td>
<td>351</td>
<td>Therostasin (4e-08)</td>
<td>AAS66770 cyst-rich cocoon protein (0.10)</td>
<td>AC164607 unannotated sequence (0.53)</td>
<td>JZ188156 Macrobdella decora similar to antistasin-like domain of cocoon protein (7e-24)</td>
</tr>
<tr>
<td>c1987_1</td>
<td>252</td>
<td>Bufrudin (1e-07)</td>
<td>P81492 Full=Hirudin-HM2; AltName: Full=Bufrudin (0.17)</td>
<td>CP003281 unannotated sequence (0.36)</td>
<td>FE383027 Daphnia pulex unannotated sequence (1.1)</td>
</tr>
<tr>
<td>c10041_1</td>
<td>213</td>
<td>Eglin C (1e-07)</td>
<td>ELU11714 unannotated sequence (7e-19)</td>
<td>CU606901 unannotated sequence (0.085)</td>
<td>EY375536 Helobdella robusta unannotated sequence (1e-11)</td>
</tr>
<tr>
<td>c2724_2</td>
<td>294</td>
<td>Saratin (2e-06)</td>
<td>2K13X Saratin (3.7)</td>
<td>AC154629 unannotated sequence (8e-04)</td>
<td>CA859348 Glomus versiforme unannotated sequence (6e-13)</td>
</tr>
<tr>
<td>c3189_1</td>
<td>369</td>
<td>Theromin (2e-06)</td>
<td>EHJ68713 unannotated sequence (0.25)</td>
<td>CP002907 unannotated sequence (0.16)</td>
<td>GW194319 Acropora palmata unannotated sequence (1.7)</td>
</tr>
<tr>
<td>c4341_1</td>
<td>678</td>
<td>LAPP (9e-05)</td>
<td>ELT87063 unannotated sequence (0.001)</td>
<td>CP003642 unannotated sequence (1.10)</td>
<td>FC893256 Citrus clementina unannotated sequence (0.27)</td>
</tr>
</tbody>
</table>

aAs calculated from the blastall pairwise alignments.
bNo protein or nucleotide sequences for manillase exist on GenBank.
cThe contig showed motifs common to the CLECT family and matched c-type lectin from Crassostrea gigas at 2e-10.
dNo nucleotide sequence for piguamerin exists on GenBank.
eAntistasin was predicted as the second best hit.
fNo nucleotide sequence for LDTI exists on GenBank.
gAn inhibitor of trypsin was predicted as the second best hit at 1e-13.
hNo nucleotide sequence for theromin exists on GenBank.
Examination of alignments

To verify BLAST-annotations, alignments for each best scoring sequence and its respective archetypal protein were examined in detail. Where the degree of conservation across the alignment was low, paying particular attention to the disulphide-bond forming cysteines (Cys), the annotation was deemed incorrect.

The cystatin alignment (Fig. 2a) showed a high degree of conservation among amino acid residues when comparing contig c3896_2 to the archetypal protein. Interestingly, both the contig and the archetypal sequence included only a single cysteine, which may hinder disulphide-bond formation. However, this cysteine showed conservation between the sequences, much like the long string of amino acid residues toward the C-terminus: QVVAGTNFVKV. The program SignalP predicted neither the contig nor the archetypal sequence in possession of a signal peptide region.

The thrombin inhibitor alignment (Fig. 2b) showed an extremely high degree of conservation, particularly toward the N-terminus, where all eight cysteines were conserved. Two of these cysteines, however, lay within the predicted signal peptide region of both the contig and archetypal sequence, which agrees well with previous findings for hirudin (three disulphide-bonds; Chatrenet & Chang 1993; Min et al. 2010); this result suggests that the unidentified thrombinhibitor represents a hirudin ortholog. The longest string of identical residues (FVVVFVAVCICVTQS) occurs toward the N-terminus.

For LDTI, the best scoring hit matched the archetypal anticoagulant at 4e$^{-13}$ and the alignment (Fig. 2c) displayed a high level of conservation between the sequences downstream of the predicted signal peptide region, which was only predicted for the contig. In addition, the six cysteines were all conserved. The best reciprocal BLASTx hit for the contig matches bdellin-KL from Hirudo nipponia WHITMAN 1886 at 4e$^{-10}$ and the best BLASTn hit matched the aspartic-rich protein aspolin at 9e$^{-19}$ (Table 2); unfortunately, no nucleotide sequence is available for LDTI in GenBank. Because of the significant hit against aspolin by the LDTI-matching contig, we also considered the alignment of those sequences (Fig. 2d). Three different factors, displayed by the alignment, conclusively showed that the contig is not part of the aspolin family of proteins: (i) the degree of conservation across the length of the alignment was much lower than that of the LDTI alignment; (ii) as opposed to the six cysteine

---

Fig. 1. Results from the Blast2GO analysis, showing the proportional allocation of the sequenced transcripts to various classification categories. The arrangement from top to bottom of the categories to the right corresponds to decreasing representation among the sequenced transcripts such that the topmost categories are the most represented.
residues possessed by the contig, only a single non-conserved cysteine was present in the aspolin sequence and this lay within a predicted signal peptide region; and (iii) most of the shared amino acid positions between the sequences lay within the predicted signal peptide region. In light of these factors, the contig appears to be correctly annotated for LDTI.

Fig. 2. MAFFT-based amino acid alignments of sequences derived from *Haemadipsa interrupta* and their respective archetypal anticoagulants. Green boxes denote conserved cysteine residues, red boxes denote the predicted signal peptide region and blue shadings represent conservation of residues between the sequences. a. putative cystatin (HT8UG4B01_c3896_2) together with AAN28679 from *Theromyzon tessulatum* (MÜLLER 1774); b. putative hirudin (HT8UG4B01_c7044_1) together with CAA79672 from *Haemadipsa sylvestris*; c. putative leech derived tryptase inhibitor (HT8UG4B01_c3999_2) together with AAB33769 from *Hirudo medicinalis*; and d. the translation of the best scoring reciprocal BLASTn nt match (AB189965) from *Cyprinus carpio* LINNAEUS 1758 (Teleostei: Cyprinidae) against the leech derived tryptase inhibitor sequence (see text for details).
The eglin c alignment (Fig. S1) displayed a high degree of conservation across the molecule. In contrast to findings that eglin c is devoid of cysteines (Min et al. 2010; Kvist et al. 2013), the best scoring eglin c contig herein showed a single cysteine at position 43. No signal peptides were predicted.

For c-type lectin, the alignment (Fig. S2) between the contig and the archetypal protein showed a rather high degree of conservation with eight conserved cysteines shared between the sequences; three additional cysteines were present in the archetypal sequence derived from the salivary EST library of *M. decora*. Both the contig and archetypal sequence contained a predicted signal peptide region at the N-terminus.

The antiplatelet alignment (Fig. S3) contained the best hits for saratin and LAPP as well as both archetypal anticoagulants. The alignment showed a high degree of conservation primarily in the mid domain of the molecule, between codon positions 85–169, whereas the conservation decreased toward the termini. Closer examination of the alignment revealed less conservation between the best scoring putative saratin contig and the archetypal anticoagulants when compared to the best scoring putative LAPP contig. This was especially true in terms of cysteine conservation; all six known cysteines of saratin/LAPP (Gronwald et al. 2008; Kvist et al. 2011a) were conserved within the putative LAPP contig, whereas only one was present in the putative saratin contig, suggesting that the annotation of the latter is incorrect. A signal peptide region was only predicted for the archetypal LAPP sequence.

The best scoring putative manillase contig aligned well with the archetypal sequence and showed a high degree of conservation, in particular toward the C-terminus, where the longest string of identical residues is SKMRLLFDLNAE (Fig. S4). The archetypal sequence did not include cysteines; however, a single cysteine was present in the contig (position 61). No signal peptide regions were predicted for either sequence.

For His-rich protein, the archetypal sequence was highly expressed in the salivary transcriptome of *M. decora* (Min et al. 2010), but its function is still unknown. Although the archetypal sequence included two mid domain insertions when compared to the contig (codon positions 31–37 and 44–52), the degree of conservation across shared amino acid positions was high (Fig. S5). The single cysteine was conserved and predicted signal peptide regions were present in both sequences.

The degree of amino acid conservation was high in the first domain of the ficolin alignment (Fig. S6) and two cysteine residues shared identical positions between the sequences; an additional cysteine was present in the archetypal sequence. In the second domain, the contig included a large insertion, which was lacking in the archetypal sequence. Judging from the levels of shared similarity, however, the contig appears to be correctly annotated for ficolin. No signal peptide regions were predicted.

The bdellin alignment (Fig. S7) also displayed high degree of conservation between the sequences across their length. This included conservation of six cysteines that are thought to be involved in formation of three disulphide bonds (Min et al. 2010; Kvist et al. 2013). In accordance with Min et al. (2010) and Kvist et al. (2013), the present study also found two proline (Pro) residues (positions 19 and 24), confirming that this unexpected finding in past studies was indeed correct (bdellin was long thought to be devoid of proline residues [Fritz et al. 1971]). The contig included a signal peptide region, suggesting that it is secreted.

In the alignment of the Kazal-type serine protease inhibitors (Fig. S8), the degree of conservation was high at the 3'-end and lower in the 5'-end. The archetypal anticoagulant, represented by an EST from *M. decora*, possessed 12 cysteines, but the contig only possessed six. Whereas the six cysteines showed conservation between the sequences, the putative annotation based on the BLAST search is still questionable.

The anti-factor Xa alignment (Fig. S9) included both matching contigs and archetypal sequences for the following anticoagulants: antistasin, guamerin, piguamerin, ghilanten, and therostasin. The number of fully, or nearly fully, conserved cysteines across the alignment (n=20) agreed well with previous findings in leech antistasin (n=22; Kvist et al. 2013). Only 10 of these occurred outside of any predicted signal peptide region; except for contig c5389_2, which matched therostasin at 4e⁻⁸, the cysteines were fully conserved in all sequences. After closer examination, *H. interrupta* appears to possess sequences with significant similarity to antistasin, guamerin, piguamerin and ghilanten, whereas the therostasin match was too disparate to confidently infer annotation. Indeed, the best reciprocal BLASTn hit (7e⁻²⁴) for contig c5389_2 against GenBank EST was a sequence from *M. decora* annotated for a protein similar to an antistasin-like domain of cysteine-rich cocoon proteins. Some cocoon proteins are paralogous to antistasin family proteins in leeches (Mason et al.
2004), and this is probably the reason that contig c5389_2 also matched therostasin, but at a much higher e-value.

Conservation in the bufrudin alignment (Fig. S10) was largely centered on the six conserved cysteine residues that occurred outside of the predicted signal peptide region (predicted for both the contig and archetypal sequence). Bufrudin, a thrombin inhibitor, is similar in length and cysteine distribution to hirudin, and the best scoring hit for bufrudin also matched hirudin but at a higher e-value (3e⁻⁶ vs. 1e⁻⁷).

In contrast to the well-conserved bioactive protein sequences noted above, some alignments were not as conserved. For example, the theromin alignment (Fig. S11) contained two large insertions in the contig when compared with the archetypal anticoagulant; even in regions of shared amino acids, conservation was remarkably low. Whereas the archetypal anticoagulant contained 16 cysteine residues, the contig only included one, which was not shared with the anticoagulant. No signal peptides were predicted. Although the e-value score (2e⁻⁶) for contig c3189_1 suggests a close affinity to theromin, the alignment clearly indicates that the contig is not related to the theromin family of proteins.

The same holds true for the leukocyte elastase inhibitor alignment (Fig. S12), which also contradicted the low e-value score (1e⁻²⁷) obtained from a BLASTx analysis against our local data set. Except for the second domain of the molecule, conservation across the alignment was minimal. A signal peptide was predicted for the archetypal sequence but no cysteines (n=1 in the anticoagulant, n=5 in the contig) were shared by the sequences. Thus, details of the alignment suggest that these sequences are most likely not orthologous and that the contig does not belong to the leukocyte elastase inhibitor family.

In conclusion, sequences from the H. interrupta transcriptome matched each of cystatin, hirudin, LDTI, eglin c, c-type lectin, LAPP, manillase, His-rich protein, ficolin, bdellin, Kazal-type serine protease inhibitor, antistasin, guamerin, piguamerin, and bufrudin with low (well scoring) e-values, identical or near identical best reciprocal BLASTx and/or BLASTn hits, and high conservation across the alignments. These three factors suggest that annotations for the contigs are correct. In addition, regions indicative of signal peptides were predicted for contigs that matched hirudin, LDTI, c-type lectin, His-rich protein, bdellin, antistasin, guamerin, piguamerin, and bufrudin, supporting the hypothesis that these proteins are actively secreted and used during bloodfeeding by the leech.

Other highly represented transcripts

The connectin titin, originally isolated from the hemichordate Saccoglossus kowalevskii (AGASSIZ 1873), was not only the most frequent hit among the assembled contigs, but was also present in three different forms among the ten most frequent hits (best hit=4e⁻³⁸, see Table S1). Whereas the remaining seven hits were largely unannotated sequences (hypothetical proteins) from various organisms, two matched sequences annotated as collagen alpha 2 (IV) at 1e⁻¹⁷. In fact, among the 50 most frequent hits matching an annotated sequence, titin, various collagen forms, filamin (3e⁻⁹⁶), twitchin-like proteins (3e⁻²⁵) and connectin (4e⁻²⁰) dominate; this prevalence is only interrupted by isolated hits matching tumor differentially expressed protein (9e⁻¹⁷), calponin (2e⁻⁹¹), protein LET-2 (2e⁻²⁸), glycine-rich secreted cement protein (2e⁻³⁴), zinc finger protein (1e⁻⁵⁰), and granulins (3e⁻²⁸) (Table S1).

A surprisingly high number of hits were recovered that matched sequences derived from other blood-feeding organisms, such as the yellow fever mosquito (Aedes aegypti [LINNAEUS 1758]), southern house mosquito (Culex quinquefasciatus [SAY 1823]), and the malaria-bearing mosquito Anopheles gambiae GILES 1902. The vast majority of these matches were against unannotated sequences, but some were of potential interest to anticoagulation activities, such as calpain (Cong et al. 1989) at an e-value of 3e⁻²⁶ and vitellogenic carboxypeptidase (Ribeiro 2003) at 9e⁻¹⁶.

In addition to these select matches against eukaryotic taxa, several matches were found against sequences derived from bacteria, predominately alphaproteobacteria (Agrobacterium spp., Rhizobium spp.; see Tables S1, S2). The three highest scoring bacterial BLASTx hits were against sequences annotated as aconitate hydratase from Agrobacterium tumefaciens Conn 1942 (e-value of 0), proteins from the porin subfamily from Rhizobium leguminosarum (FRANK 1879) (1e⁻¹⁵⁵), and outer membrane protein OmpA from an Agrobacterium spp. (1e⁻¹⁵⁴). Interestingly, several significant hits were found against bacteria-derived protein catalysts, specifically those regulating the biosynthesis of riboflavin (Richer et al. 1997; Ludwig et al. 1987). These include bifunctional riboflavin deaminase-reductase (at 4e⁻³⁵), riboflavin synthase subunit beta (at 4e⁻³⁸), and 6,7-dimethyl-8-ribityllumazine synthase (at 1e⁻⁵⁶), all from A. tumefaciens.
Phylogenies

As mentioned above, phylogenetic analyses were limited to those bioactive protein families for which three or more amino acid sequences with probable orthology were available, which excluded cystatin, thrombin, His-rich proteins, and Kazal-type serine protease inhibitors. The terminology used to describe the unrooted trees shown here follows that of Wilkinson et al. (2007), such that a “clan” is the equivalent to a “monophyletic group” in a rooted tree and “adjacent group” is equivalent to “sister-group” in a rooted tree. Summary statistics for the trees are presented in Table 3. By and large, the clans in the trees do not conform to species clusters, i.e., there is little species fidelity among the select proteins.

In the hirudin tree (Fig. 3), contig c7044_1 from *H. interrupta* forms a clan with three archetypal sequences; one annotated as haemadin and two as thrombin inhibitors, and all from *H. sylvestris* BLANCHARD 1894. Moreover, contig c1987_1, which scored well against both hirudin and bufrudin in the BLASTx and BLASTn searches, places as the adjacent group to a larger set of sequences including those mentioned above as well as a putative hirudin ortholog derived from a *M. decora* EST library. The contrast between the position of this sequence and the archetypal bufrudin sequence indicates that both *H. interrupta* sequences are hirudin (or possibly haemadin) orthologs.

The antiplatelet tree (Fig. S15), including putative saratin and LAPP orthologs as well as the archetypal sequences, shows a high amount of structure in terms of both taxonomic fidelity and separation of the different proteins. The two main clans in the tree represent the saratin orthologs and the LAPP orthologs, respectively. The two contigs from *H. interrupta* (*c2724_2* and *c4341_1*) form a clan adjacent to the archetypal sequences of LAPP derived from glossiphioid leeches. This solidifies the results from the examination of the alignments, in that no saratin orthologs are present in the *H. interrupta* data set. Instead, both contigs seem to represent LAPP putative orthologs.

Unfortunately, the manillase tree (Fig. S16) offers little help in assigning orthology as the tree is completely unresolved. What this might suggest, however, is that the included sequences are all part of the same radiation, with considerable subsequent differentiation. To this end, the putative manillase

---

Table 3. Summary statistics for the aligned amino acid data sets for each bioactive protein family and resulting phylogenetic trees.

<table>
<thead>
<tr>
<th>Bioactive protein</th>
<th>No. aligned sites</th>
<th>No. of most parsimonious trees</th>
<th>Length</th>
<th>CI</th>
<th>RI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hirudin</td>
<td>136</td>
<td>2</td>
<td>206</td>
<td>0.971</td>
<td>0.937</td>
</tr>
<tr>
<td>LDTI</td>
<td>225</td>
<td>2</td>
<td>101</td>
<td>0.990</td>
<td>0.857</td>
</tr>
<tr>
<td>Eglin c</td>
<td>124</td>
<td>2</td>
<td>177</td>
<td>0.977</td>
<td>0.826</td>
</tr>
<tr>
<td>C-type lectin</td>
<td>193</td>
<td>1</td>
<td>459</td>
<td>0.950</td>
<td>0.763</td>
</tr>
<tr>
<td>Saratin/LAPP</td>
<td>241</td>
<td>6</td>
<td>899</td>
<td>0.811</td>
<td>0.833</td>
</tr>
</tbody>
</table>

CI, consistency index; RI, retention index.
Three main clans are presented in the ficolin tree (Fig. S17), the most interesting of which includes both the archetypal sequence from *Hi. medicinalis* and the *H. interrupta* contig (c1477_1). The branches within this clan are also somewhat shorter than the remaining branches, which suggests that the *H. interrupta* contig indeed represents a ficolin ortholog.

There is little taxonomic structure in the bdellin tree (Fig. S18), with sequences from the various taxa placing in several different clans. Together with a sequence from *M. decora*, the *H. interrupta* contig (c2073_1) places as the adjacent group to a cluster of *Hirudo*-derived sequences, including the archetypal sequence. Judging from this placement, the contig seems to represent a true bdellin ortholog.

The tree from the factor Xa inhibitor data set (Fig. S19) is almost completely unresolved, offering no substantial information on the putative orthology determinations of the *H. interrupta* contigs (c5389_2, 5961_1, 6218_1, c14138_1 and c3664_1). Much like the manillase data set, it seems as though there is either too low phylogenetic signal among the polymorphic sites or that there are too few polymorphic sites in the data set. Regardless of this, it seems as though there is no evidence for paralogy or non-homology of each of the putative factor Xa inhibitors derived from the *H. interrupta* transcriptome (therostasin, antistasin, guamerin, piguamerin, and ghilanten).

The tree generated from the leukocyte elastase inhibitor data set (Fig. S20) is completely unresolved, which is somewhat surprising as the alignment of the *H. interrupta*-derived contig (c729_5) and the archetypal sequence from *Hi. medicinalis* suggested a non-affiliation between the sequences. Thus, it remains unclear whether or not contig c729_5 is a true leukocyte elastase ortholog.
Discussion

Pyrosequencing of the salivary gland transcriptome of the terrestrial leech *Haemadipsa interrupta* generated sequences that show strong resemblance to a large diversity of proteins related to anticoagulation. Transcripts significantly matched 20 different known bioactive proteins, many of which employ different antagonistic pathways; 13 of these had identical or near identical best reciprocal BLAST hits when targeted against GenBank, six matched unannotated sequences best, and only one matched a different annotation when reciprocally BLASTed. For several of the contigs, the best reciprocal BLAST hits matched at insignificant e-values, typically orders of magnitude higher than the best hit against the local data set. This can be attributed to at least two factors. First, some proteins in the locally compiled database lack a nucleotide, EST, and/or protein counterpart in GenBank (this impelled our performing BLAST searches against a variety of databases). Second, as the size of the BLAST target database increases, the probability of recovering a match by chance alone increases, thereby increasing the e-values for each hit (Altschul et al. 1997); the same is true when the nucleotide length of the query is low, and this may also lead to inflated e-values.

Subsequent to BLAST-based analyses, detailed examination of alignments further supported our hypothesis that cystatin, hirudin, LDTI, eglin c, c-type lectin, LAPP, manillase, His-rich protein, ficolin, bdellin, Kazal-type serine protease inhibitors, antistasin, guamerin, piguamerin, and bufrudin are possessed by *H. interrupta*, and that roughly half of these include a signal peptide region (Table 2). The apparent lack of secretory signal peptides for some of the sequences can be set in context when reviewing previous work on leech salivary proteins. Our imposed restriction on *getorf* to return only ORFs starting with an initiation codon (Met) meant that it is possible that some of the three domains of the signal peptide region (Von Heijne 1990) have not been recovered, such that SignalP does not predict a signal peptide. Indeed, both Min et al. (2010) and Kvist et al. (2013) found several predicted secretory signal peptides in sequences without an initial methionine. To this end, the non-sampled full length of the proteins investigated in the present study might still be endowed with a transient extension beyond the methionine that would allow for the mature protein to be translocated across the cell membrane.

The final security tier for our orthology inferences was to examine the phylogenetic trees derived from the different protein data sets. Although some of the trees offered little information due to their low resolution, others verified the BLAST-based orthology. In particular, the trees generated from the hirudin, LDTI, antiplatelet, and ficolin data sets solidify our orthology determinations. It can be concluded from these analyses that our alignment-based hypotheses of orthology for almost all of the bioactive proteins were correct, as they were also confirmed by the phylogenetic analyses, except for the sequence best matching bufrudin, which is likely a hirudin ortholog.

Below we discuss the significance of matches to known bioactive proteins and frame them in an evolutionary context by comparing the relative placement of their antagonistic pathways in a recent,
The evolution of bloodfeeding in leeches

Although the evolution of hematophagy in leeches has been studied extensively, hypotheses related to this subject have been somewhat contradictory depending on the type and amount of data included in the study. Sawyer (1986) first suggested that the most recent common ancestor of extant leeches did not employ hematophagy and that rhynchobdellid (proboscis-bearing) leeches and hirudiniforms evolved their bloodfeeding behavior independently. Siddall & Burreson (1995) reconstructed one of the earliest molecular phylogenies for leeches (based on mitochondrial COI) and found that their results did not directly conflict with that of Sawyer (1986); the authors noted that bloodfeeding seems to have evolved independently at least twice in leeches—once along the branch leading to Rhynchobdellida, and another time along the branch leading to Hirudiniformes (Siddall & Burreson 1995, 1996). Independent origins of this life-history strategy in leeches are further supported by differences in the anatomical features through which bloodfeeding is carried out. Rhynchobdellid leeches possess an eversible proboscis that pierces the skin of the prey, whereas arhynchobdellid leeches are armed with three jaws and create an incision wound using a sawing motion (Siddall & Burreson 1996). Regardless of this fundamental discrepancy, more contemporary and data-rich studies have converged on the opinion that bloodfeeding is a plesiotypic strategy among leeches. For example, using nuclear 18S rRNA and mitochondrial 12S rRNA, Trontelj et al. (1999) recovered the families Piscicolidae and Glossiphoniidae as evolving on two subsequent branches. The most parsimonious character optimization of bloodfeeding on that tree involved a hematophagous most recent common ancestor of leeches and three subsequent losses in predaceous and liquidosomatophagous (those that feed on haemolymphal fluids) glossiphonids, erpobdelliforms and haemophil hirudiniforms. Two independent lines of evidence support the hypothesis that the ancestral leech was bloodfeeding. First, Kvist et al. (2011a) showed that leech antiplatelet proteins, which inhibit von Willebrand factor-mediated activation of subendothelial platelets, are present in the genome of the non-bloodfeeding glossiphoniid leech Helobdella robusta. Shanksland et al. 1992. As Glossiphoniidae is commonly recovered at the base of the leech phylogeny (Siddall & Burreson 1998; Apakupakul et al. 1999; Light & Siddall 1999; Siddall et al. 2001; but see Trontelj et al. 1999), it is likely that He. robusta possesses anticoagulants simply by virtue of their being passed down from a common ancestor. Second, Hirabayashi et al. (1998), Joo et al. (2009), and Lee et al. (2010) each isolated and characterized anticoagulation-related proteins from lumbricid ciliates (galactose-binding lectin from Lumbricus terrestris LINNAEUS 1758 and factor Xa inhibiting eisenstatin from Eisenia andrei BOUCHÉ 1972). Because leeches are derived oligochaetous annelids (Rousset et al. 2007; Zrzavý et al. 2009; Struck et al. 2011; Kvist & Siddall 2013), it can be hypothesized that the archetypal leech possessed some “oligochaete”-inherited anticoagulants similar in composition to those of modern leeches. The last point, as it relates to the evolution of hematophagy in leeches, has largely been neglected in the literature, yet may have a fundamental impact on our understanding of it.

Recent advancements in the generation of data concerning the presence of anticoagulants in various leeches (Min et al. 2010; Kvist et al. 2011a, 2013; Siddall et al. 2011) have increased our understanding of the evolution of bloodfeeding, inasmuch as anticoagulants are directly related to bloodfeeding capacity. These data also lend themselves well to a timely discussion on the presence and diversity of anticoagulants across leech phylogeny. Figure 5 illustrates our current understanding of leech phylogeny (Min et al. 2010; Siddall et al. 2011), as well as our contemporary knowledge of the distribution of the antagonistic pathways used by several bioactive proteins across the diversity of leeches. The figure bears witness to a growing interest in salivary transcriptomics in the larger and perhaps more charismatic leech families; Glossiphonidae, Macrobdellidae, Haemadipsidae, and Hirudinidae have been heavily studied when compared to other bloodfeeding families such as Piscicolidae, Praobdellidae, and Xerobdellidae. Whereas the absence of bioactive salivary proteins related to bloodfeeding in Ameribdellidae, Gastrocomobdellidae, Salipidae, Haemopidae, Semiscoleceidae, and to a large extent, Erpobdellidae may be reflective of their predatory lifestyle, their absence in the bloodfeeding Piscicolidae, Praobdellidae, and Xerobdellidae may simply reflect a lack of data for these groups. One of the more general insights provided by the tree is that most of the pathways used by bioactive proteins possessed by representatives of arhynchobdellid families (all families except Glossiphonidae and Piscicolidae) are also present in Glossiphonidae, the latter of which represents the earliest diverging lineage of leeches. Therefore, it seems likely that these
Fig. 5. Composite metaphylogeny of leeches (adapted from Min et al. 2010; Siddall et al. 2011) with a matrix representation of our current understanding of the distribution of antagonistic pathways used by bioactive salivary proteins in the various taxa. Each terminal represents a single specimen and corresponding familial affinity is noted. In the matrix, check marks indicate a known presence of salivary proteins with specific antagonistic pathways, and question marks indicate either that the salivary protein repertoire of the taxon is unknown or that studies failed to recover proteins with those specific antagonistic pathways. Ame, Americobdellidae; Cyl, Cylicothelliidae; Erp, Erpobdellidae; Gas, Gastrostomobdellidae; Glo, Glossiphoniidae; Hae, Haemadipsidae; Hao, Haemopidae; Hir, Hirudinidae; Mac, Macrobdellidae; Oli, Oligochaetous clitellates; Pis, Piscicolidae; Pra, Praobdellidae; Sal, Salifidae; Sem, Semiscoleidesidae; Xer, Xerobdellidae. The photograph at bottom right shows living individuals of Haemadipsa interrupta.
proteins are present in arhynchobdellid taxa simply by virtue of their presence in a common glossiphoniid or piscicolid ancestor, or both (Fig. 5). Insofar as some of these pathways refer to those used by known and well-characterized anticoagulants, this further supports the hypothesis of a bloodfeeding ancestral leech.

Based on our current knowledge of bioactive protein pathway distribution, it appears that some glycoprotein IIb/IIIa inhibitors may not be present in arhynchobdellid taxa outside of Macrobdellidae. For example, prior studies have failed to obtain decorsin-like sequences from representatives of Hirudinidae (Kvist et al. 2013), and the present study shows that the same holds true for a haemadipsid species. However, ornatin, a paralog of decorsin, has been isolated from at least one glossiphoniid leech, Placobdella ornata (Verrill 1872) (Mazur et al. 1993), which could suggest its subsequent loss in Hirudinidae and Haemadipsidae and further evolution in Macrobdellidae.

Anticoagulant-associated proteins and bacterial hits

A number of ORFs significantly matched sequences that were annotated as bacterial; the predominant match was to alphaproteobacteria (Agrobacterium spp., Rhizobium spp.; see Tables S1, S2). While this may be the result of external contamination during dissection or RNA extraction, great care was taken to avoid such contamination, and we hypothesize that these sequences originated from an internal infection. Several families of leeches host endosymbiotic alpha- and gammag proteobacteria (Kikuchi & Fukatsu 2002; Siddall et al. 2004, 2011; Perkins et al. 2005; Graff et al. 2006; Kvist et al. 2011b). It has even been hypothesized that bacteria supply leeches with nutrients and essential vitamins that are deficient in their diet. Vertebrate blood is notoriously low in vitamin B and, in the present study, we found several high-scoring hits against proteins that play a direct role in the biosynthesis of riboflavin (vitamin B2). These findings provide yet another line of evidence in support of a symbiotic relationship between leeches and bacteria that likely is beneficial to the host (Kvist et al. 2011b). Prior studies have indicated that leeches of the family Glossiphoniidae enjoy symbiotic relations with mycetome-associated gamma- and alphaproteobacteria, whereas hirudiniforms are most often associated strictly with gammag proteobacterial symbionts (Graff et al. 2006; Siddall et al. 2011). Fluorescent in situ hybridization studies, using class-specific probes, would reveal whether or not the bacterial contigs found here stem from an external contamination or internal infection; such studies are a necessity before any robust conclusions can be drawn about the association between haemadipsid leeches and alphaproteobacterial endosymbionts.

As mentioned above, several other bioactive salivary proteins of interest were recovered in the H. interrupta salivary transcriptome. The majority of BLAST annotations were to a variety of different forms of anticoagulant-related proteins and all of them scored significant e-values. The specific structure and function of these proteins remain unclear, but it seems likely that they are related to blood-feeding and anticoagulation. The most significant hits are briefly discussed below in the context of their potential contribution to anticoagulation.

At a range of e-values (the lowest scoring at 1e−16), a total of 98 transcripts matched sequences that were annotated as various forms of cathepsin. Cathepsin is thought to play a role in inhibiting the coagulation cascade by cleaving and thus inactivating several coagulation factors, including fibrinogen. The function of cathepsin is similar to that of hirudin, in that it cleaves, and thereby potentially modulates, the function of the thrombin receptor.

Disintegrins are small yet potent proteins that inhibit platelet aggregation and integrin-dependent cell adhesion (McLane et al. 2004). Various disintegrins have been isolated from snake venom where they also act as metalloproteinases (Kini & Evans 1992; Jia et al. 1996). A number of contigs (n=24) from the H. interrupta salivary transcriptome matched various forms of disintegrins, the best hit scoring an e-value of 2e−22.

In addition, several transcripts matched sequences annotated as a 29 kDa galactose-binding lectin (at 1e−57 and above). As with disintegrins, galactose-binding lectins are frequently recovered from snake venoms where they are involved in anticoagulation and antiplatelet aggregation (Xu et al. 1999). Interestingly, the best scoring transcript matched a galactose-binding lectin isolated from the lumbricid earthworm, L. terrestris, supporting the aforementioned hypothesis that ancestors of the archetypal leech did indeed possess anticoagulants.

In response to parasitic infections, host mast cells release a granule-associated enzyme to augment wound healing and defend against pathogens (Reverter et al. 1998; Prussin & Metcalfe 2003). Leech carboxypeptidase inhibitors (LCIs) are thought to block a host’s defenses by inhibiting proteases of mast cells (Reverter et al. 1998). A sequence annotated as LCI was retrieved as the best match for several transcripts from H. interrupta, the best hit scoring at 6e−18.

Invertebrate Biology
vol. 133, no. 1, March 2014

Kvist, Brugler, Goh, Giribet, & Siddall
Anticoagulant repertoire versus diet

In contrast to the dietary preferences of many freshwater leeches (poikilotherm blood), terrestrial leeches are typically less fastidious in their prey choice. Considering their aptitude for locating suitable sources of blood in tropical jungles, one can surmise that these leeches must be able to feed on anything that crosses their path: mammals, reptiles, and amphibians alike (e.g., Fogden & Proctor 1985; Schnell et al. 2012). A strategy such as this would necessitate a wide array of anticoagulants to counteract the wide variety of coagulation factors present in different prey. Specimens of H. interrupta are notoriously aggressive and feed on a variety of prey (Seagrace & Salsman 1946; Stammers 1950; Bhatia 1975; Sawyer 1986; Kutschera et al. 2007). Prior to this study, it was unknown whether the diversity of anticoagulants in the repertoire of a single leech was correlated with dietary choices. The results presented herein provide a first line of evidence in support of a positive relationship between the diversity of anticoagulants and variety of prey items, as H. interrupta possesses one of the most inclusive anticoagulant repertoires recovered in a single species of leech. This may suggest that the broad anticoagulant repertoire of H. interrupta has evolved in response to the dietary strategy of the species, and that other terrestrial leeches, particularly those with fewer dietary restrictions, will probably contain a similarly broad repertoire. This is further supported by the fact that H. interrupta secretes a wide variety of proteins that may have anticoagulatory properties, even beyond the obvious anticoagulants (see above). Min et al. (2010) and Kvist et al. (2013) did not recover all of these proteins from EST libraries of the aquatic Hirudo verbana, Macrob melodella decora, and Altolimnatis fenestrata. However, differential gene expression during various life-stages of the leeches may alter the potential of finding anticoagulants; therefore, this result should be viewed with appropriate caution. That is, it is possible that each of the aforementioned leech species does produce a wider array of anticoagulants than found in previous studies but that they were not expressed at the time of RNA extraction. In terms of bioactive protein pathways, the repertoire of H. interrupta does not possess unique or significantly broader components when compared to aquatic leeches.

Conclusions

Through molecular characterization and phylogenetic inference, we demonstrate that the terrestrial, Southeast Asian haemadipsid leech H. interrupta possesses a diverse set of salivary proteins putatively involved in anticoagulation. Furthermore, several additional bioactive proteins were found in the transcriptome, most of which are lacking complete information regarding structure and function in leeches. This suggests that leeches may contain more anticoagulation factors than previously predicted. Forthcoming salivary transcriptomic studies should focus on increasing taxon sampling in an effort to map the presence of anticoagulants across the taxonomic diversity of leeches and thus gain a better understanding of the evolution of salivary proteins in leeches. Whereas H. interrupta possesses a diverse anticoagulant repertoire, the antagonistic pathways of putative anticoagulants do not fundamentally differ from those of their aquatic counterparts. As putative anticoagulants have also been found in oligochaetous clitellates, a complete picture of anticoagulant evolution cannot be obtained without specifically investigating and integrating data from these key taxa.

Acknowledgments. This research was funded by generous support from the Wenner-Gren Foundations, the Sixten Gemzéus Foundation, Helge Ax:son Johnson’s Foundation, and the Olle Engkvist Byggmästare Foundation to S.K. Fieldwork was funded by the Museum of Zoology, University of Malaya Grant 2011. We thank the Willi Hennig Society for making TNT freely available. Editor Michael Hart and two anonymous reviewers provided important comments that greatly improved the manuscript. The authors declare no conflict of interest.

References


Keller PM, Schultz LD, Condra C, Karczewski J, & Connolly TM 1992. An inhibitor of collagen-stimulated platelet activation from the salivary glands of the Hae-
Anticoagulant diversity in Haemadipsa interrupta


Supporting information

Additional Supporting information may be found in the online version of this article.

Fig. S1. MAFFT-based amino acid alignment of putative egln-c from Haemadipsa interrupta (HT8UG4B01_c10041_1) against the known sequence of the anticoagulant protein (0905140A) from an unknown hirudinid. Blue shadings represent conservation of residues between the sequences.

Fig. S2. MAFFT-based amino acid alignment of putative lectoxin-like c-type lectin from Haemadipsa interrupta (HT8UG4B01_c14817_4) against a sequence of the putative bioactive protein (Macrobdella2B10strict38) from Macrobdella decora. Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and blue shadings represent conservation of residues between the sequences.

Fig. S3. MAFFT-based amino acid alignment of putative saratin and LAPP from Haemadipsa interrupta (HT8UG4B01_c2724_2 and HT8UG4B01_rep_c4341_1, respectively) against the known sequences of the anticoagulants (2K13_X and Q01747) from Haementeria officinalis DE Filippi 1849. Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and blue shadings represent conservation of residues between the sequences.

Fig. S4. MAFFT-based amino acid alignment of putative manillase from Haemadipsa interrupta (HT8UG4B01_c1130_2) against the known sequence of the bioactive protein (Patent no. 2006 US 7.049.124 B1) from Hirudinaria manillensis (LESSON 1842). Blue shadings represent conservation of residues between the sequences. Fig. S5. MAFFT-based amino acid alignment of putative His-rich protein from Haemadipsa interrupta (HT8UG4B01_rep_c3900_1) against a sequence of the putative bioactive protein (Macrobdella14F06strict17) from Macrobdella decora. Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and blue shadings represent conservation of residues between the sequences.

Fig. S6. MAFFT-based amino acid alignment of putative ficolin from Haemadipsa interrupta (HT8UG4B01_c1477_1) against a sequence of the putative bioactive protein (Macrobdella10A03strict686) from Macrobdella decora. Green boxes denote conserved cysteines and blue shadings represent conservation of residues between the sequences.

Fig. S7. MAFFT-based amino acid alignment of putative bdellin from Haemadipsa interrupta (HT8UG4B01_c2073_1) against the known sequence of the anticoagulant (P09865) from Hirudo medicinalis. Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and blue shadings represent conservation of residues between the sequences.

Fig. S8. MAFFT-based amino acid alignment of putative Kazal-type serine proteinase inhibitor from Haemadipsa interrupta (HT8UG4B01_rep_c4204_4) against a sequence of the putative bioactive protein (Macrobdella14D11strict26) from Macrobdella decora. Green boxes denote conserved cysteines and blue shadings represent conservation of residues between the sequences.

Fig. S9. MAFFT-based amino acid alignment of putative factor Xa inhibitors (HT8UG4B01_c14138_1 guamerin, HT8UG4B01_rep_c6218_1 piguamerin, HT8UG4B01_rep_c5961_1 antistasin, HT8UG4B01_c3664_1 ghilanten and HT8UG4B01_rep_c5389_2 therostasin) from Haemadipsa interrupta against the known sequences of the anticoagulants (AAD09442 guamerin, P81499 piguamerin, AAA29193 antistasin, P16242 ghilanten, and Q9NBW4 therostasin) from Hirudo nipponia (guamerin and piguamerin), Haementeria officinalis, Haementeria ghilianii DE Filippi 1849, and Theromyzon tessulatum, respectively. Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and intensity of blue shadings corresponds to BLOSUM62 conservation of residues.

Fig. S10. MAFFT-based amino acid alignment of putative bufrudin from Haemadipsa interrupta (HT8UG4B01_c1987_1) against the known sequence of the anticoagulant (P81492) from Poecilobdella manillensis (LESSON 1842). Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and blue shadings represent conservation of residues between the sequences.
Fig. S11. MAFFT-based amino acid alignment of putative theromin from *Haemadipsa interrupta* (HT8UG4B01_c3189_1) against the known sequence of the anticoagulant (P82354) from *Theromyzon tessulatum*. Blue shadings represent conservation of residues between the sequences.

Fig. S12. MAFFT-based amino acid alignment of putative leukocyte elastase inhibitor from *Haemadipsa interrupta* (HT8UG4B01_c729_5) against a sequence of the putative bioactive protein (MacrobdellaElastaseInhib) from *Macrobdella decora*. Red boxes denote the predicted signal peptide region, green boxes denote conserved cysteines, and blue shadings represent conservation of residues between the sequences.

Fig. S13. Unrooted strict consensus of two equally parsimonious trees recovered from the analysis of the eglin c data set (see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Fig. S14. Unrooted single most parsimonious tree recovered from the analysis of the c-type lectin data set (see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Fig. S15. Unrooted strict consensus of six equally parsimonious trees recovered from the analysis of the antiplatelet (saratin and LAPP) data set (see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Fig. S16. Unrooted strict consensus of 19 equally parsimonious trees recovered from the analysis of the manillase data set (see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Fig. S17. Unrooted single most parsimonious tree recovered from the analysis of the ficolin data set (see also Table 3). Branch lengths are proportional to change.

Fig. S18. Unrooted single most parsimonious tree recovered from the analysis of the bdellin data set (see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Fig. S19. Unrooted strict consensus of 28 equally parsimonious trees recovered from the analysis of the factor Xa inhibitor data set (antistasin, therostasin, guamerin, piguamerin, ghilanten; see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Fig. S20. Unrooted strict consensus of three equally parsimonious trees recovered from the analysis of the leukocyte elastase inhibitor data set (see also Table 3). When appropriate, GenBank accession numbers follow taxon names. Branch lengths are proportional to change.

Table S1. Results from the BLASTx searches against NCBI, using the contigs as queries.

Table S2. Results from the BLASTn searches against NCBI, using the contigs as queries.