Draft genome sequences of *Hirudo medicinalis* and salivary transcriptome of three closely related medicinal leeches.

Vladislav V. Babenko[1], Oleg V. Podgorny[1,2], Valentin A. Manuvera[1,3], Artem S. Kasianov[3,4], Alexander I. Manolov[1], Ekaterina N. Grafskaia[1,3], Dmitriy A. Shirokov[1], Alexey S. Kurdyumov[1], Dmitriy V. Vinogradov[5,6], Anastasia S. Nikitina[1,3], Sergey I. Kovalchuk[1,7], Nickolay A. Anikanov[1,7], Ivan O. Butenko[1], Olga V. Pobeguts[1], Daria S. Matushkina[1], Daria V. Rakitina[1], Elena S. Kostryukova[1], Victor G. Zgoda[8], Isolda P. Baskova[9], Vladimir M. Trukhan[10], Mikhail S. Gelfand[5,6,11,12], Vadim M. Govorun[1,3], Helgi B. Schiöth[10, 13] Vassili N. Lazarev [1,3]

[1]Federal Research and Clinical Centre of Physical-Chemical Medicine of Federal Medical Biological Agency, 1a Malaya Pirogovskaya Str., Moscow 119435, Russia

[2]Koltzov Institute of Developmental Biology, Russian Academy of Sciences, 26 Vavilov str., Moscow 119334, Russia

[3]Moscow Institute of Physics and Technology, 9 Institutskiy per., Dolgoprudny, Moscow Region 141700, Russia

[4]Vavilov Institute of General Genetics, Russian Academy of Sciences, 3 Gubkina str., Moscow 119991, Russia

[5]A.A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, 19 Bol'shoi Karetnyi per., Moscow 127051, Russia

[6]Skolkovo Institute of Science and Technology, 3 Nobelya Ulitsa str., Moscow 121205, Russia

[7]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, 16/10 Miklukho-Maklaya str., Moscow 117997, Russia

[8]V.N. Orekhovich Research Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, 10 Pogodinskaja str., Moscow 119832, Russia

[9]Faculty of Biology, Lomonosov Moscow State University, 1-12 Leninskie Gory, Moscow 119991, Russia.

[10]I.M. Sechenov First Moscow State Medical University of the Ministry of Healthcare of the Russian Federation (Sechenovskiy University), Trubetskaya str., 8-2, Moscow 119991, Russia

[11]Faculty of Computer Science, National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow 101000, Russia

[12]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, 1-73 Leninskie Gory, Moscow 119991, Russia

[13]Functional Pharmacology, Department of Neuroscience, Uppsala University, Uppsala, Sweden.

Abstract

Salivary cell secretion (SCS) plays a critical role in blood feeding by medicinal leeches, making them of use for certain medical purposes even today. We annotated the *Hirudo medicinalis* genome and performed RNA-seq on salivary cells isolated from three closely related leech species, *H. medicinalis, Hirudo orientalis*, and *Hirudo verbana*. Differential expression analysis verified by proteomics identified salivary cell-specific genes, many of which encode previously unknown salivary components. However, the genes encoding known anticoagulants were not differentially expressed in the salivary cells. The function-related analysis of the unique salivary cell genes enabled an update of the concept of interactions between salivary proteins and components of haemostasis. Thus, our

study provides one of the most comprehensive knowledge of the genetic fundamentals of the blood-sucking lifestyle in leeches.

Introduction

The genome sequencing of haematophagous animals and transcriptional profiling of their salivary glands has attracted considerable attention in recent years because many haematophagous species transmit various infectious diseases caused by viruses, bacteria, protozoa, and helminths. The elucidation of the genetic mechanisms that allow haematophagous species to act as vectors of pathogenic organisms [1–5] is of great importance for public health care, veterinary medicine, and microbiology. Opposite to other hematophagous species, blood-sucking leeches, belonging to the subclass *Hirudinea* (true leeches) of the phylum *Anelidae*, attract interest in regard to the identification of novel bioactive compounds. Blood-sucking leeches and, in particular, medicinal leeches, have been used for bloodletting to treat diverse ailments since ancient times [6]. Although the use of live leeches to treat human diseases is not encouraged by current medicine because of the high risk of an undesired outcome, hirudotherapy is still indicated in certain medical conditions. In particular, application of medicinal leeches improves tissue drainage after replantation when the common surgical correction of venous congestion fails or is unfeasible [7]. In these cases, hirudotherapy frequently provides beneficial effects because the leech feeding apparatus has evolved to promote the finely tuned inhibition of haemostasis and blood coagulation [8, 9]. The composition of leech saliva has been shown to play a key role in this inhibition [9, 10].

In medicinal leeches, which belong to the group of so-called *jawed leeches*, saliva is secreted by unicellular salivary glands that reside in the anterior part of the body and are interspersed between the muscle fibres that connect the jaws with the body wall. Each salivary cell extends a single duct from the cell body to the jaw, and the duct ends in a tiny opening between the calcified teeth of the jaw. Medicinal leeches incise the host skin at the feeding site by their jaws and release the salivary cell secretion (SCS) into the wound [9]. During one act of blood meal, medicinal leeches partially or completely empty their salivary gland cells [11]. However, salivary gland cells replenish their reservoirs with the content of SCS within seven days after blood meal, making a leech to be got ready for another act of blood feeding [12]. In addition to its known inhibitory effect on haemostasis and blood coagulation, SCS suppresses inflammation, exhibits analgesic effects, possesses antimicrobial activity, and alters vasodilatory responses, enhancing local blood circulation to facilitate leech feeding. Moreover, some SCS components are thought to preserve the blood from rapid degradation after ingestion.

Although transcriptional analysis of the salivary glands of jawless and jawed leeches was attempted [13–16] and some SCS components have been characterized and their respective targets in hosts have been identified [9], the repertoire of the bioactive saliva components remains largely unknown. Elucidating the SCS composition in medicinal leeches is the key to understanding (i) the molecular mechanisms underlying the orchestrated interaction between the leech SCS and the components of haemostasis, (ii) evolution of the blood feeding lifestyle in leeches, and (iii) genetics of hematophagy. Identification of the unknown bioactive SCS components will facilitate the development of novel pharmacological compounds for treating impaired peripheral blood circulation, venous congestion or microbial infections.

In the current study, we performed sequencing, genome assembly and annotation *H. medicinalis* genome as well as transcriptional profiling of the salivary cells followed by proteomic validation of SCSs of three medicinal leeches, *Hirudo medicinalis, Hirudo orientalis*, and *Hirudo verbana*. This study aims to provide a comprehensive map of the genetically encoded components of blood meal-related genes in leeches.

Results

Genome assembly and annotation

To assemble the *H. medicinalis* genome, we extracted DNA from an adult leech. Before being processed, the leach was maintained without feeding for at least two months. We created a set of three shotgun libraries to perform sequencing by using three different platforms (**Supplementary Table 1**). All read datasets were combined, and a single assembly was created by SPAdes [17]. The resulting assembly contained 168,624 contigs with an N50 contig length of 12.9 kb (**Supplementary Table 2**).

Preliminary analysis (contigs Blast) revealed the presence of bacterial sequences in the resulting assembly. Therefore, we conducted binning to discriminate the leach contigs (a leech bin). We built a distribution of contigs according to their GC abundance, tetranucleotide frequencies, and read coverage. To increase the binning accuracy, the read coverage was determined by combining the DNA reads with the reads corresponding to a combined transcriptome of *H. medicinalis* (see below). The discrimination of the eukaryotic and prokaryotic contigs is illustrated in **Fig. 1A/B, Supplementary Table 3** and **Supplementary Data 2.** Additionally, we selected the mitochondrial contigs to assemble the leech mitochondrial genome [18].

The eukaryotic contigs underwent a scaffolding procedure using paired reads. Scaffolds were generated using Illumina paired-end and mate-pair read datasets by SSPACE [19]. After scaffolding, the assembly consisted of 14,042 sequences with an N50 scaffold length of 98 kb (**Supplementary Tables 4 and 5**). The total length of the genome draft was estimated to be 187.5 Mbp, which corresponds to 85% of the theoretical size of the leech genome (**Supplementary Table 6**). A total of 14,596 protein coding genes were predicted.

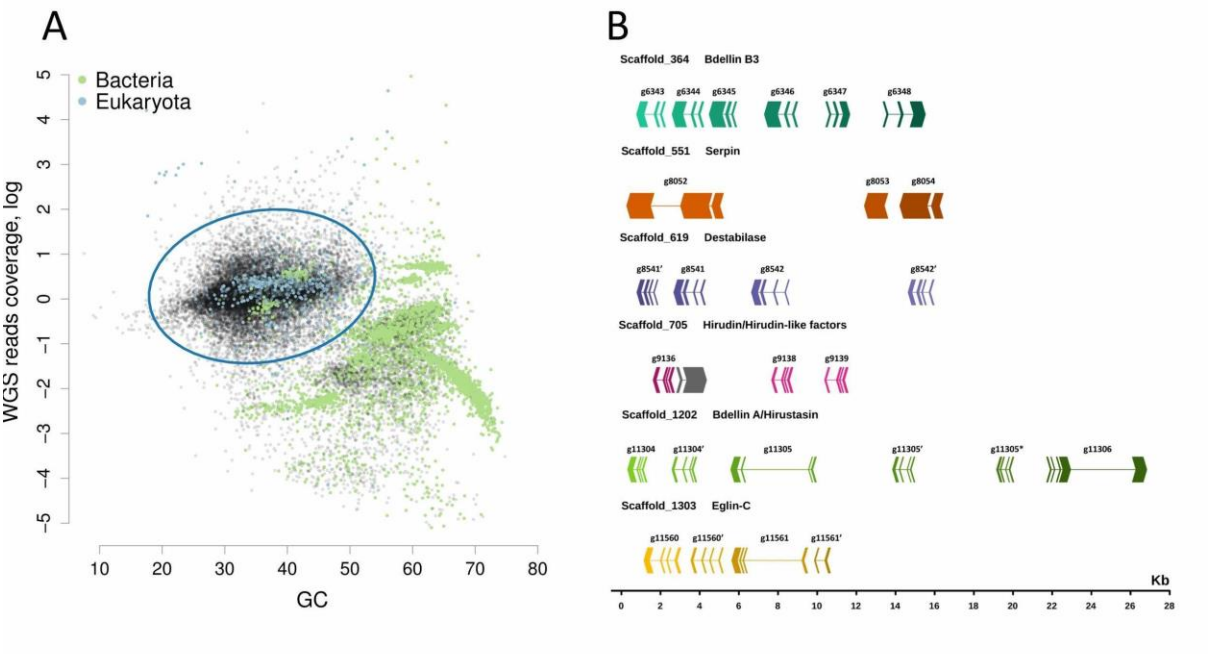

Fig. 1. The *H. medicinalis* genome binning. **(A)**. 2D-plot showing the contig distribution in coordinates of GC content and coverage by a combination of reads obtained by Ion Proton and Illumina. Contigs are indicated by dots, and the taxonomic affiliation of contigs at the domain level is encoded by colour (green – *Bacteria*, blue – *Eukarya*, black – no assignment). The taxonomic affiliation was determined by direct Blastn search against the National Center for Biotechnology Information (NCBI)

nt database. The 3D plot showing the contig distribution in coordinates of GC content, read coverage (Proton and Illumina), and host cDNA read coverage is presented in **Supplementary Data 1**. **(B)** *H. medicinalis* genome contains clusters of blood meal-related genes. The graph shows the exon-intron structure of genes and arrangement of gene clusters in scaffolds on a general scale. The exon arrows indicate the direction of transcription (gray - unknown gene).

Also, we identified new homologs of genes encoding known anticoagulants or blood meal-related proteins. The multiple amino acid alignments for each of these protein families (**Supplementary Figs. 1,2**) Based on the genome sequence data and using known protein sequences, we determined the organization of these genes (**Supplementary Table 7, Fig1 B.**). Positions and lengths of exons and introns were predicted using the respective cDNA and protein sequences as references. In some cases, genes are localized in common scaffolds and form tandems or clusters **Fig1 B.**

mRNA-seq, transcriptome assembly and annotation

To obtain tissue-specific mRNA samples from three medicinal leech species, *H. medicinalis*, *H. verbana*, and *H. orientalis*, we isolated salivary cells and muscles from the cryosections of the anterior body parts using laser microdissection (**Fig. 2A**). Then, we constructed two cDNA libraries with and without normalization for each mRNA sample using the oligo-dT primer and sequenced them on the Ion Torrent PGM (**Supplementary Table 8**). Four read datasets corresponding to the constructed cDNA libraries were used for the *de novo* assembly of a combined transcriptome for each medicinal leech species using the Trinity RNA assembler [20] (**Supplementary Table 9**). We used the combined transcriptomes to map non-normalized tissue-specific reads. Read mapping was necessary to perform consecutive differential expression analysis.

Gene Ontology (GO) analysis of the detected transcripts was performed using Blast2GO [21] and BlastX. The 'nr' database served as a reference database. GO analysis demonstrated that all three medicinal leech species had similar transcript distributions across GO categories (**Supplementary Figure 3**). The taxonomy distribution of the closest Blast hits also was similar (**Supplementary Figure 4**). The majority of the identified transcripts were found to match two species of *Annelida*: 59.8% to *H. robusta* and 10.7% to *C. teleta*. This analysis also confirmed the absence of contamination by non-leech transcripts.

The prediction of coding regions (or open read frames, ORFs) and annotation of transcriptomic data were carried out using Transdecoder and Trinotate. ORFs were translated using the BlastP algorithm, and the protein sequences were annotated by EuKaryotic Orthologous Groups (KOG) classification using the eggNOG database [22] (**Supplementary Figure 5**). The KOG classification revealed that all three medicinal leech species have similar transcript distributions across KOG categories. All three medicinal leech species were also found to share the vast majority of their orthologous clusters (**Supplementary Figure 6**).

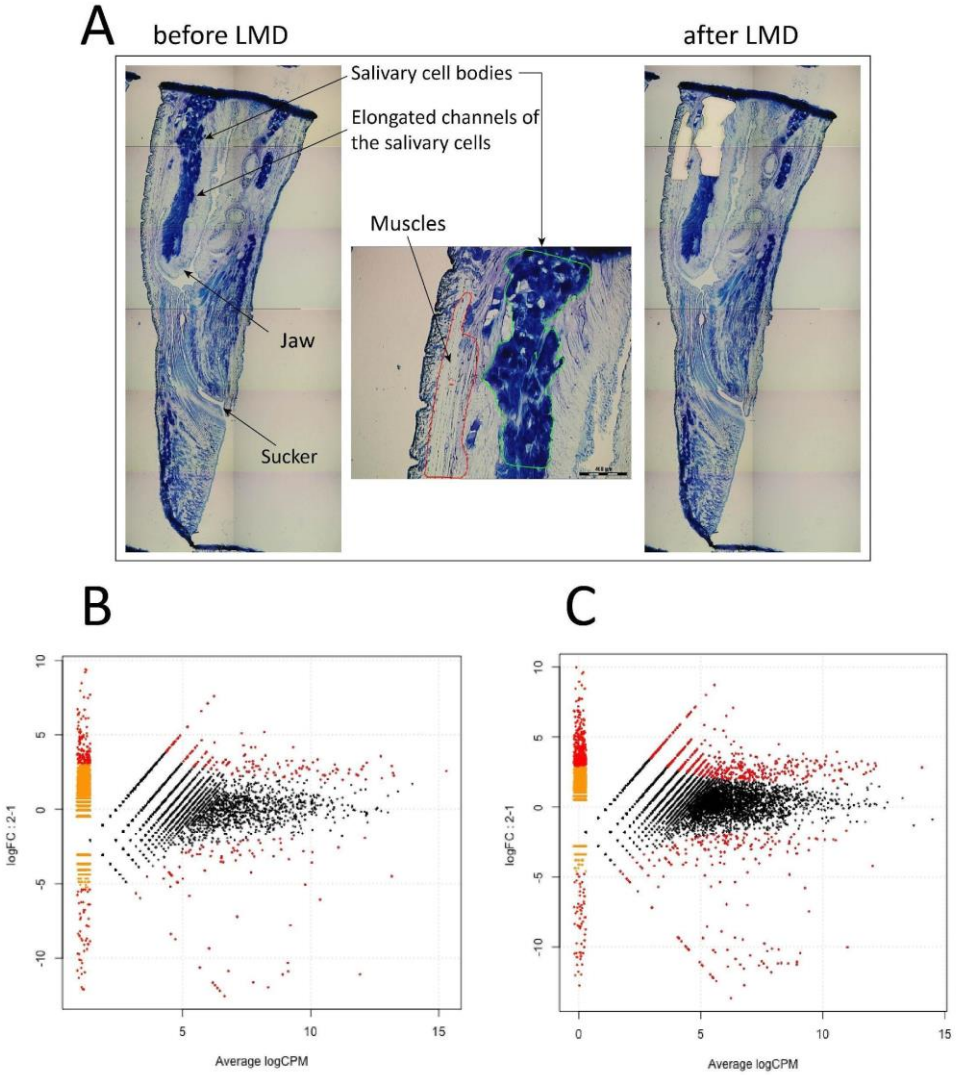

Fig. 2 Differential expression analysis of salivary cells. **(A)** Isolation of salivary cells and muscles by laser microdissection. MA plots of differentially expressed genes in the salivary cells and muscles of *H. medicinalis* for the *de novo* assembled transcriptome **(B)** and the genome model **(C)**.

Differential expression analysis

To estimate the relative expression levels of the transcripts identified in the salivary cells and muscles and to identify transcripts unique to the salivary cells, we mapped the tissue-specific cDNA reads without normalization against the combined transcriptome of each medicinal leech species. We also mapped the tissue-specific cDNA reads of *H. medicinalis* against its genome assembly. Differentially expressed genes were detected according to a recent protocol [23]. To identify genes that are differentially expressed in the salivary cells and muscles, an individual MA plot was constructed for each medicinal leech species using its combined transcriptome (**Fig. 2B, Supplementary Figure 7**). An additional MA was constructed for *H. medicinalis* using its genome assembly (**Fig. 2C**). Genes with a q-value (FDR) < 0.05 were considered to be differentially expressed.

We identified 102, 174, and 72 differentially expressed transcripts in the salivary cells of *H. medicinalis, H. orientalis*, and *H. verbana*, respectively. Because the three are closely related medicinal leech species, the protein sequences of the differentially expressed transcripts were

grouped into orthologous clusters to simplify the subsequent functional analysis. We identified 25 differentially expressed, orthologous clusters shared by three leech species and 44 orthologous clusters shared by at least two leech species (**Fig. 3, Supplementary Tables 10-11**). The majority of sequences in the identified orthologous clusters correspond to hypothetical proteins annotated in the genome of *H. robusta*. Analysis of conserved domains in the identified orthologous clusters allowed the determination of sequences belonging to known protein families.

We also analysed the differentially expressed genes of *H. medicinalis* using its genome assembly. The cDNA reads for the salivary cells, muscles, and neural tissue [24] (reads were obtained from the Sequence Read Archive (SRA)) were mapped onto the genome assembly. For the neural tissue, we used a read dataset for ganglion 2 because of its localization in the preoral segments. Differential expression analysis identified 42 genes unique to the salivary cells of *H. medicinalis* (**Supplementary Table 12**).

Proteomics of salivary cell secretion

For proteomic analysis, we collected SCSs from three medicinal leech species, *H. medicinalis, H. orientalis*, and *H. verbana*, which were maintained without feeding for at least two months. The SCSs were collected according to a previously reported method [25] with some modifications (see Methods).

The sample preparation method is critical to the resultant repertoire of the identified proteins because the SCS consists of both low- and high-molecular-weight components [9] and contains proteinase inhibitors, glycoprotein complexes, and lipids. The latter may form complexes with proteins [26]. Therefore, we combined several sample preparation methods and several mass spectrometry techniques to cover the broadest repertoire of the SCS proteins. Proteomic datasets obtained by different sample preparation methods and mass spectrometry techniques were combined to create a final list of the identified proteins for each medicinal leech species.

We identified 189, 86, 344 proteins in the SCSs of *H. medicinalis, H. orientalis*, and *H. verbana*, respectively and grouped them into orthologous clusters as described above. All three medicinal leech species were found to share 39 orthologous clusters, and 50 orthologous clusters were shared by at least two species (**Fig. 3, Supplementary Table 13**). Combination of the transcriptomic and proteomic data revealed 25 orthologous clusters of genes unique to the salivary cells (**Supplementary Table 11**). A list of individual components of the leech SCS is given in **Fig. 3**. Surprisingly, the expression of genes encoding known SCS anticoagulants and blood meal-related proteins did not differ in salivary cells and in muscles. To validate this finding, we examined the expression of saratin, eglin C, bdellins, hirustasin, destabilase, metallocarboxypeptidase inhibitor, apyrase, and angiotensin converting enzyme (ACE) by the real time PCR of additional, independent tissue-specific cDNA libraries constructed for salivary cells and muscles. The real-time PCR results (data not shown) confirmed this finding. This indicates that genes encoding anticoagulants and blood meal-related proteins are involved not only in the blood feeding, but contribute to other, yet unknown physiological functions.

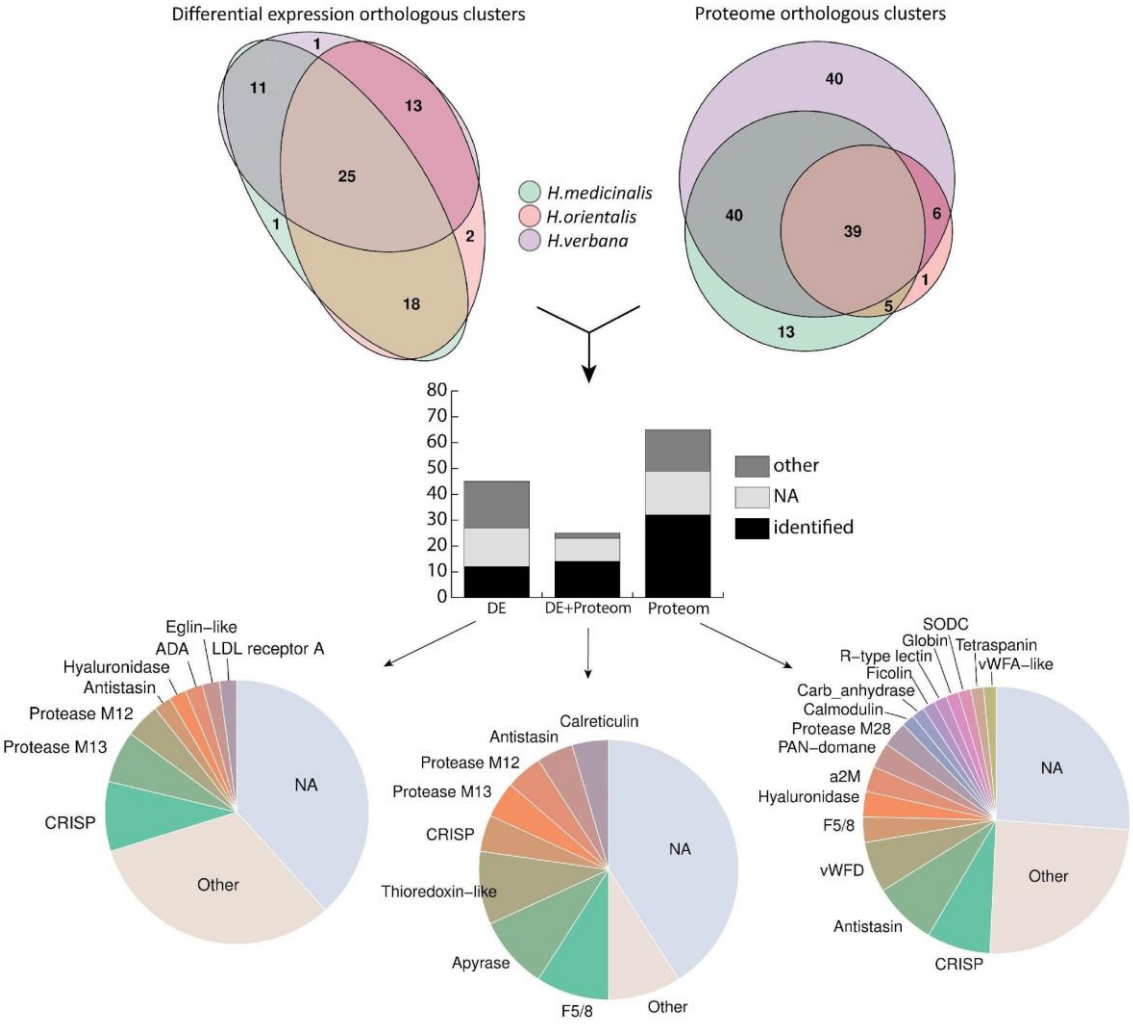

Fig. 3. Summary of the identified SCS components. The Venn diagrams in the upper panel show the numbers of ortholog clusters identified by differential expression (DE) and proteomic (Prot) analyses across three medicinal leech species. The histogram in the middle panel features the numbers of orthologous clusters identified by the differential expression analysis, proteomic analysis or a combination thereof (DE+Prot). Each bar consists of ortholog clusters identified as known blood feeding-related components (identified), other known proteins (other), and unknown proteins (NA). The pie charts in the lower panel illustrate the abundance of the individual SCS components identified by the differential expression analysis, proteomic analysis or their combination. For details, see **Supplementary Tables 11, 12, and 14**.

Below, we characterize SCS components classified into functional groups and describe their possible roles in the hemostasis. The sequences of proteins and their alignment are presented in **Supplementary Figs. 8-23.**

Enzymes

Proteases: The results of this study show that metalloproteases of the M12, M13, and M28 families are the major enzymatic components of the SCS. The M12B (ADAM/reprolysin) peptidases are a large family of disintegrin-like metalloproteinases that have a broad range of functions and are involved in many physiological processes [27]. These enzymes are often found in snake venoms while the transcripts are observed in sialotranscriptomes of various hematophagous species [28–30]. In haemostasis, secreted proteases of the M12 family can participate in the inhibition of platelet

adhesion [31, 32] and in clot softening due to the degradation of fibrinogen. These proteins exhibit metal-dependent proteolytic activity against extracellular matrix proteins (gelatine, fibrinogen, fibronectin), thereby affecting the regulation of inflammation and immune responses.

In mammals, proteases of the M13 family are involved in the formation and development of the cardiovascular system and in the regulation of neuropeptides in the central nervous system [33]. One of their most important functions is the activation of biologically active peptides, particularly peptides involved in the regulation of blood pressure (angiotensin and bradykinin). In mammals, ACE is an important component of the renin angiotensin system (RAS). ACE is expressed in the sialotranscriptomes of the leech (*Theromyzon tessulatum*), the cone snail (*Conidae*), the vampire snail (*Colubraria reticulata*), and dipteran species (*Diptera*) [34, 35].

The identified sequences of M28 family exopeptidases belong to the Q-type carboxypeptidases, also known as lysosomal dipeptidases or plasma glutamate carboxypeptidase (PGCP). These peptidases were shown to be involved in the regulation of the metabolism of secreted peptides in the blood plasma and the central nervous system in mammals [36]. These enzymes appear to serve to deactivate certain signalling peptides in the blood and are components of haemoglobinolytic systems in haematophagous parasites, playing the role of digestive exopeptidases [37]. Notably, leech salivary gland secretions contain carboxypeptidase inhibitors, which presumably prevent the untimely digestion of the blood meal by other types of peptidases [9, 38].

Superoxide dismutase (EC 1.15.1.1): We identified sequences of secreted superoxide dismutase family (SODC, Cu/Zn type) enzymes. This family of metalloproteins is mainly typical of eukaryotes and is involved in free radical inactivation, which retards oxidative processes. In the blood, superoxide dismutase catalyses the conversion of superoxide into molecular oxygen and hydrogen peroxide and prevents the formation of peroxynitrite and hydroxyl radicals [39]. Interestingly, peroxynitrite may suppress haemostatic function by the nitration of key procoagulants [39, 40], while hydrogen peroxide is a key signalling molecule involved in the regulation of many processes (coagulation, thrombosis, fibrinolysis, angiogenesis, and proliferation). In ticks, SODC is presumed to participate in regulating the colonization of the intestinal tract by bacteria, including causative agents of diseases [41]. In SCSs, SODC appears to exhibit an antibacterial effect along with other proteins of the innate immune system and prevents unwanted blood oxidation during feeding and digestion. Notably, haem-containing compounds and free iron are involved in the formation of free radicals and the provocation of oxidative stress [42].

Carbonic anhydrase (EC 4.2.1.1): This enzyme is a key component of the bicarbonate buffer system and is involved in the regulation of pH values in the blood, the digestive tract, and other tissues [43, 44]. In haematophagous animals, this enzyme can maintain optimal conditions for the digestion of a blood meal [45, 46]. Carbonic anhydrase appears to cause a local increase in acidosis at the bite site, decreasing the activity of blood coagulation factors.

Hyaluronidase (EC 3.2.1.35): These enzymes are common in the proteomic and transcriptomic data of haematophagous and venomous animals. The salivary secretions of different leech species are known to contain hyaluronidase (heparinase, orgelase) [47]. In the proteome and transcriptome, we found three clusters containing a domain of the glycosyl hydrolase family 79 (O-glycosyl hydrolases). This family includes heparinases, which play an important role in connective tissues. In venoms and salivary gland secretions, these enzymes catalyse the hydrolysis of hyaluronic acid, resulting in the loss of structural integrity of the extracellular matrix and thereby facilitating the penetration of anticoagulants and other active molecules deeper into the tissues [48]. In addition, the low-molecular-weight heparin produced by cleavage suppresses and inhibits blood coagulation [49].

Apyrase (EC 3.6.1.5): Apyrases are nucleotidases involved in the enzymatic degradation of ATP and

ADP to AMP. Secreted apyrases and 5'-nucleases are well-known and well-characterized components of the salivary gland secretions of venomous and haematophagous animals, including leeches [10]. Apyrases are anticoagulants because they remove ADP, an important inducer of platelet aggregation at sites of tissue injury [50].

Adenosine/AMP deaminase (EC:3.5.4.4): Adenosine deaminase (ADA) catalyses the hydrolytic deamination of adenosine to form inosine. Adenosine deaminases are well studied and have been found in the saliva of various blood-sucking insects [51]. ADA is also found in the salivary gland secretion of the vampire snail *C. reticulata*, which belongs to *Spiralia,* as well as leeches [52]. ADA is thought to play an important role in the removal of adenosine because of its involvement in pain perception processes [53].

Proteinase inhibitors

Antistasins: We identified sequences corresponding to proteinase inhibitor I15 (leech antistasin) **Fig.4**. Proteins of this family are commonly found in blood-sucking leeches and play a key role in the inhibition of blood coagulation. Their main targets are serine proteases participating in haemostasis, such as the factor Xa, kallikrein, plasmin, and thrombin [9]. Decorsin, an antistasin from *Macrobdella decora,* was demonstrated to inhibit platelet aggregation [54], and gigastasin from the giant Amazon leech (*Hementaria ghilianii*) was recently reported to potently inhibit complement C1 [55]. Antistasin from *Hementeria officinalis* is the closest homologue of the sequences identified in our study.

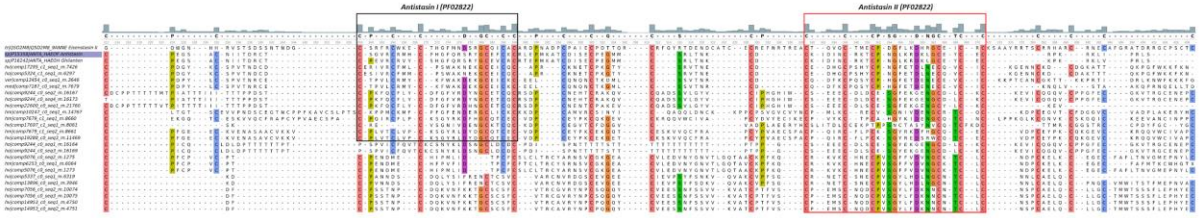

Fig. 4. Multiple sequences alignment of Antistasin-like transcripts with dual domain antistasin-type protease inhibitors from leeches' Antistasin (*Haementeria officinalis*, P15358), Ghilantein (*Haementeria ghilianii*, P16242) and Eisenstasin II from earthworm (*Eisenia andrei,* Q5D2M8*)*. The boxes indicates two antistasins domane. Alignment is generated by MUSCLE algorithm, residues are colored according to ClustalX colour scheme, conserved amino acids are colorized by conservation level (threshold > 50%). Reference sequence are marked purple.

CAP/CRISP: The cysteine-rich secretory protein/antigen 5/pathogenesis-related 1 proteins (CAP) superfamily includes numerous protein families, particularly cysteine-rich secretory protein (CRISP) **Fig 5A**. They are commonly found in the venoms of snakes and other reptiles, and most of them are toxinsy [56, 57]. In some investigations, CRISPs from haematophagous species were thought to be involved in haemostasis (HP1). The identified sequences show similarity to protein sequences from the haematophagous parasitic nematode *Ancylostoma caninum* (hookworm), such as the potassium channel blocker AcK1 [58] and the possible platelet aggregation inhibitor HPI [59], as well as to the snake toxins triflin (*Protobothrops flavoviridis*) and natrin-1 (*Naja atra*) [60, 61]. Among the differentially expressed genes, we identified sequences with a new "Cys-rich" motif **Fig 5B**. This group of proteins is characterized by the presence of a signal peptide and two cysteine patterns CX{5,14}CX{7}CX{8}CC{2}C and CX{7,17}CX{9}CX{8}CC{2}C.

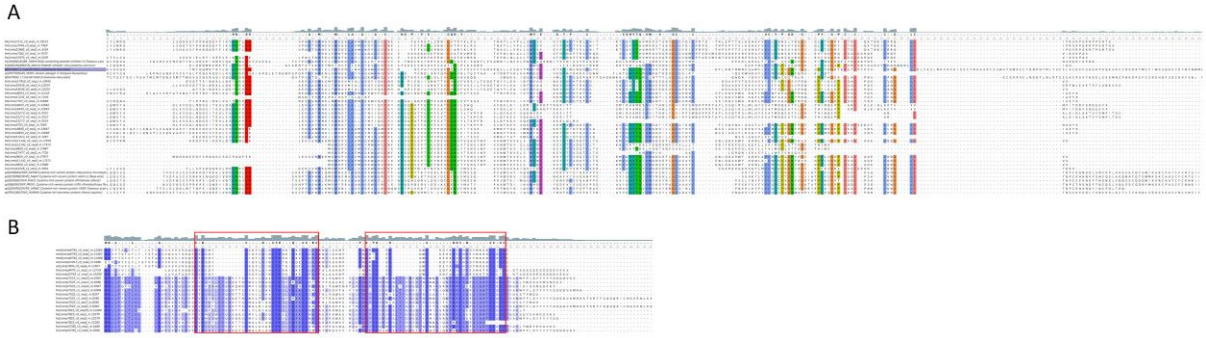

Fig. 5. (A) Alignment of CRISP domains with diverse CAP/CRISP proteins. Putative platelet inhibitors from (Ancylostoma caninum, Q962V9), (Tabanus yao, C8YJ99), CAP domain containing proteins from Vampire Snail (Cumia reticulata, QBH70087.1; QBH70092.1) and reptile Cystein-rich venom proteins triflin (Protobothrops flavoviridis), natrin-2 (Naja atra) and other. Alignment is generated by MUSCLE algorithm, residues are colored according to ClustalX colour scheme, conserved amino acids are colorized by conservation level (threshold > 50%). Reference sequence are marked purple. (B) Alignment of new "Cys-rich" domains. The boxes indicates two cysteine patterns, amino acids are colorized by percentage Identity coloring scheme.

Eglin-like: Eglins are small cysteine-free proteins that belong to the I13 family of serine proteinase inhibitors [62]. Eglins from leeches have inhibitory activity against neutrophil elastases and cathepsins G and also to participate in the protection of the crop contents from untimely proteolysis [9]. Of note, sequences identified in the present study have low homology to the classical eglin from leech **Fig.6A**.

Cystatin: We identified a cystatin sequence only in the proteome of *H. verbana*. Cystatins are small protein inhibitors of cysteine proteases (cathepsins B, H, C, L, S) [63] and are often found in the sialotranscriptomes of various ticks [64]. In ticks, cystatins play an important role in processes related to immune response, the regulation of endogenous cysteine proteases involved in the digestion of blood and haem detoxification [65]. The nematode *Nippostrongylus brasiliensis* utilizes cystatins to evade the host immune system [66].

PAN domain: This domain is present in numerous proteins, including the blood proteins plasminogen and coagulation factor XI [67]. The PAN/apple domain of plasma prekallikrein is known to mediate its binding to high-molecular-weight kininogen, and the PAN/apple domain of the factor XI binds to the factors XIIa and IX, platelets, kininogen, and heparin [68]. The salivary gland secretion of the leech *H. officinalis* was found to contain the leech antiplatelet protein (LAPP), which has a PAN domain and is involved in haemostasis. This protein exhibits affinity for collagens I, III, and IV and thereby inhibits collagen-mediated platelet adhesion [69].

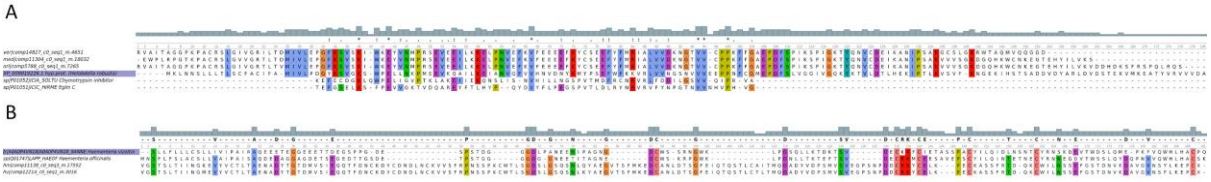

Fig. 6. (A) Amino acid sequences alignment of Eglin-like transcripts with Eglin (*Hirudo medicinalis*, P01051), hypothetical protein (*Helobdella robusta,* xp_009019226.1) and chymotrypsin inhibitor homolog from Potato *(Solanum tuberosum*, P01052). Alignment is generated by MUSCLE algorithm, residues are colored according to ClustalX colour scheme. Identical and conserved residues indicated respectively by asterisk, period and colon. (B) Alignment of PAN domains with leech anti-platelet protein (*Haementeria officinalis*, Q01747) and putative anti-platelet-like protein (*Haementeria vizottoi*, A0A0P4VN18). Conserved amino acids are colorized by conservation level (threshold > 75%). Reference sequence are marked purple.

Alpha-2-macroglobulin (α2M): The highly conserved, multifunctional α2M is involved in the inhibition of a broad range of proteases (serine, cysteine, aspartic, and metalloproteases), interacts with cytokines and hormones, and plays a role in zinc and copper chelation [70]. It can act as a plasmin inhibitor, thereby inhibiting fibrinolysis, but in some cases, it inhibits coagulation by inactivating thrombin and kallikrein [71]. This protein is believed not only to be involved in leech immune processes but also to be an important component of the salivary gland secretion that enhances anticoagulation processes.

Molecules involved in adhesion

Ficolin: Ficolins are a component of the innate immune system and trigger a lectin-dependent pathway of complement activation [72]. In invertebrates, ficolins are involved in the recognition of bacterial cell wall components [73]. The fibrinogen-like domain is present in proteins with affinity for erythrocytes, *e.g.*, tachylectin-5A (TL5A). TL5A exhibits strong haemagglutinating and antibacterial activity in the presence of Ca^{2+} ions [74]. In reptile venoms, ficolin-like proteins, ryncolin [75] (from *Cerberus rynchops*) and veficolin-1 (UniProt: E2IYB3) (from *Varanus komodoensis)*, are presumed to trigger platelet aggregation and blood coagulation.

F5/8 type C domain**:** A number of identified sequences contain one or several discoidin motifs (DS), known as the F5/8 type C domain. This domain is present in numerous transmembrane and extracellular proteins, *e.g.*, neuropilins, neurexin IV, and discoidin domain receptor proteins, and in proteins involved in haemostasis, such as the coagulation factors V and VIII [76]. The DS domain plays an important role in the binding of various ligand molecules, including phospholipids and carbohydrates [77]. Due to these features, DS-containing proteins are actively involved in cell adhesion, migration, and the proliferation and activation of signalling cascades [78]. Leech DS domain-containing proteins appear to act as lectins with high affinity to galactose and may be components of the innate immune system of the leech. In addition, they can bind to collagen or phosphatidylserine on the surface of platelets and the endothelium [79] and thus, by competitive inhibition, impair interactions between haemostatic factors.

Low-density lipoprotein receptor A family: The low-density lipoprotein receptor (LDLR) family is an important component of the blood plasma and is involved in the recognition and endocytosis of low-density lipoproteins in mammalian blood [80]. In contrast to known homologous proteins, these receptors are secretory rather than membrane proteins, and they contain four LDLR class A (cysteine-rich) repeats. Some invertebrates, including segmented worms, are hypothesized to be incapable of the synthesis of cholesterol and steroid hormones, and during feeding, leeches acquire cholesterol mainly from the blood of the host as an exogenous source [81]. We posit that this protein may be utilized by the leech for the scavenging and transportation of cholesterol-rich lipoprotein complexes.

R-type lectin: Proteins that contain the ricin-type beta-trefoil lectin domain have been found in prokaryotes and eukaryotes. In animals, R-type lectins exhibit diverse activities [82]. They are present in scavenger receptors (mannose, fucose, collagen receptors), N-acetylgalactosaminyltransferases, haemolytic toxins (CEL-III from *Cucumaria echinata*) and apoptosis-inducing cytotoxins [82, 83]. Previously, similar sequences were identified in leech transcriptomes; however, the authors assumed that this molecule has a mitochondrial localization [84]. Yet another noteworthy close homologue is the galactose-binding lectin EW29 from the earthworm *Lumbricus terrestris.* EW29 consists of two homologous domains and has been experimentally demonstrated to exhibit haemagglutinating activity [85]. As many known R-type lectins are involved in adhesion and trigger haemolysis [82], this molecule is of interest for further study.

vWFA domain: This domain is present in various plasma proteins: complement factors, integrins, and collagens VI, VII, XII, and XIV [86]. One protein identified in the leech proteome is a secreted protein that consists of four copies of the *vWFA* domain **Fig. 7**. The sequence contains several putative recognition sites: the metal ion-dependent adhesion site (MIDAS), the integrin-collagen binding site, and the glycoprotein Ib (GpIb) binding site. According to blast analysis, this domain is homologous to type VI collagen. Considering the domain organization of the protein and the presence of glycoprotein and collagen binding sites, one of the putative mechanisms of action involves binding to the surface of the endothelium or platelets, thereby preventing their interaction with collagen. This binding underlies the competitive inhibition during haemostasis (platelet scavenging) [34].

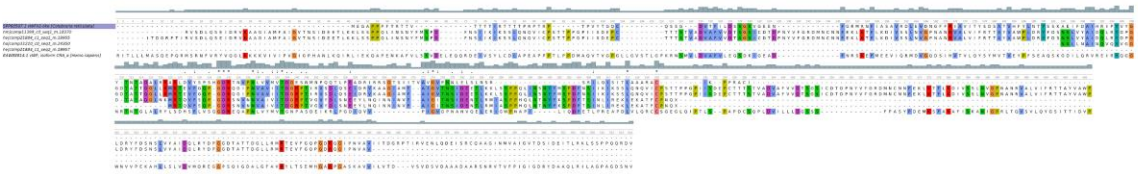

Fig. 7. Alignment of the hirudo vWFA domains with the human vWFA1 (EAW88814.1) and vWFA1-like (*Colubraria reticulata*, SPP68597.1). Alignment is generated by MUSCLE algorithm, residues are colored according to ClustalX colour scheme. Identical and conserved residues indicated respectively by asterisk, period and colon. Reference sequence are marked purple.

Discussion

Combined analysis of transcriptomic and proteomic data revealed that SCSs of three medicinal leech species, *H. medicinalis, Hirudo orientalis*, and *Hirudo verbena*, have a similar composition and to large extent contain many proteins homologous to those that are associated with either blood feeding in various hematophagous animals or have been identified in venoms of the poisonous organisms, such as M12 and M13 proteases, CRISP, Apyrase, ADA, cystatins, hyaluronidase and ficolins. However, we also detected novel salivary proteins, the roles and importance of which in blood feeding have yet to be established such as F5/8 type C domain, LDLR, PAN domain as well as proteins containing vWFD and vWFA domains. Moreover, we also found, somewhat surprisingly, that genes encoding well-characterised components of the medicinal leech SCS, such as hirudin, bdellins, eglins, saratins and destabilases were not differentially expressed in the salivary cells, suggesting that the proteins encoded by these genes support not only blood feeding but also some unknown functions in the leech.

We also identified new homologs of genes encoding known anticoagulants including antistasins and LAPP but also other blood meal-related proteins such as bdellins A and B3, eglin-C, hirustasin, destabilase, earatin/LAPP and leech DTI in the annotated genome. Interestingly, some of these genes are localized in common scaffolds and form tandems or clusters. This arrangement and the similarity of the genes suggest recent local duplications and many of them as likely to be linage specific. This unique diversity of anticoagulatant genes highlights their specific functional significance for the leech, which has led to their positive selection with subsequent expansion in the genome. It is also notable that almost all of these genes are serine protease inhibitors such as the antistasins, serpins, Kazal-Type Serine Protease Inhibitor, Potato inhibitor I family and hirudins. All of the above protein families are involved in host hemostasis or blood digestion [9, 84]. They are also the most broadly represented and diverse families in the salivary secret of leeches (**Supplementary Tables 10,11,13**).

It is possible that the initially duplication of some of these genes could be associated with dosage effect related to the amount of the gene product. During the divergence process, it is likely that some of these genes acquired increased specificity for new targets from the blood of various hosts. In the transcriptome data, we found that in most families of anticoagulants and products of blood meal related genes, the only one variant of the gene is expressed and these were the previously known sequences from the NCBI. At the same time, we also identified the expression of several variants of Bdellin B3, Lectin C-type, Destabilase, and three Elastase inhibitor genes (serpins). Gene duplication combined with positive selection is an important mechanism for the creation of new functional role for the existing genes (neofunctionalization) [87]. The functional segregation of hirudin and hirudin-like factor genes could be an example of this. The functions of hirudin are well understood, while targets of hirudin-like factors are still not known [88, 89]. Moreover, the results show that many well known blood meal-related genes are found in single-copies (without duplication) including Saratin, Leech DTI and Carboxy peptidase inhibitor. These gene families have probably very conserved functional activities in blood processing during feeding and we see no signs duplication of within these gene families.

For a long time, thrombin and factor Xa have been thought to be the main targets of components of the leech SCS. However, similar to those of other haematophagous species, the SCS of medicinal leeches were found to contain a complex mixture of molecules that inhibit both secondary and primary. Our findings lead to the suggestion that the action of the SCS components not merely on the key coagulation factors (thrombin and factor Xa) but on the different stages of haemostasis provides synergistic effects. Furthermore, functional analysis of the identified proteins demonstrated that leeches utilize ancient, highly conserved molecules as versatile affectors of the host haemostasis and immunity.

Similarity of the SCS compositions in the examined leeches with those of many other blood feeding species favours hypotheses about a convergent evolution of the saliva composition in evolutionary distant hematophagous animals as a result of having to adapt to the same targets in their hosts.

In the proteome of the SCSs, we found proteins that usually exhibit cytoplasmic or membrane localization, such as calreticulin, calmodulin, thioredoxin, chaperones, tetraspanin transcription factors, and certain ribosomal and cytoskeletal proteins. In contrast to that in jawless leeches [90], the mode of the secretion process in jawed leeches remains unknown. The presence of proteins with cytoplasmic or membrane localization appears however to be associated with apocrine secretion in the production of leech saliva.

Over all it is interesting that the SCS of the medicinal leeches have been found to contain proteins homologous to others found in a variety of the vertebrate species. These proteins are relatively conserved, and exhibit shared structural and functional features among various species. Some have been shown to be directly or indirectly involved in haemostasis in mammals. These proteins, avoiding interactions with components of the host immune system, could possibly affect the kinetics of biochemical reactions, thereby providing a synergistic effect of the SCS.

In sum, this analysis provides new insights into the role of the genome structure in the regulation of blood feeding-related gene expression and the evolutionary adaptation to the blood-sucking lifestyle. More broadly, the genome annotation performed in our study may serve as a blueprint for future experimentation on the medicinal leech as a model organism and provides a database of sequences encoding the unique bioactive leech proteins for use in developing novel pharmacological compounds.

Methods

BioProject and raw sequence data. The genome assembly was validated by the National Center for Biotechnology Information (NCBI). It was checked for adaptors, primers, gaps, and low-complexity regions. The genome assembly was approved, and the accession numbers MPNW00000000 and BioProject PRJNA257563 were assigned. All genome sequencing data were deposited in the Sequence Read Archive (SRA) with accession numbers (see **Supplementary Table 1**). Raw mRNA-seq data are available as FASTQ files and have been quality-checked and deposited in the SRA with their accession numbers (see **Supplementary Table 9**).

Biological samples. Three leech species were provided by HIRUD I.N. Ltd. (Balakovo, Saratov Region, Russia). *H. medicinalis, H. verbana,* and *H. orientalis* were collected at a pond near Volkovo, Saratov region, Russia (51°91′03″, 47°34′90″), Lake Manych, Stavropol Krai, Russia (46°01′09″, 43°48′21″), and Lake Divichi, Kura–South Caspian Drainages, Azerbaijan (41°17'40″, 49°04'13"), respectively. Leech species were confirmed by sequencing the regions of the genes encoding nuclear and mitochondrial ribosomal RNA (for details, see **Supplementary Methods**).

Genomic DNA extraction, WGS sequencing and genome assembly. Genomic DNA was extracted from a single adult leech *H. medicinalis* using the standard technique with slight modifications (for details see **Supplementary Methods**). The extracted DNA was purified (for details, see **Supplementary Methods**), and a set of three shotgun libraries was created.

Ion Proton shotgun sequencing. Pure genomic DNA (approx. 1000 ng) was fragmented to a mean size of 200-300 bp using the Covaris S220 System (Covaris, Woburn, Massachusetts, USA). Then, an Ion Xpress™ Plus Fragment Library Kit (Life Technologies) was employed to prepare a barcoded shotgun library. Emulsion PCR was performed using the One Touch system (Life Technologies). Beads were prepared using the One Touch 2 and Template Kit v2, and sequencing was performed using Ion Proton 200 Sequencing Kit v2 and the P1 Ion chip.

Ion Torrent mate-pair sequencing. A mate-pair library with 3-6 kb fragments was prepared from pure genomic DNA using Ion TrueMate Library Reagents (Life Technologies). The Ion PGM™ template OT2 400 kit (Life Technologies) was used to conduct emulsion PCR. Sequencing was performed by the Ion Torrent PGM (Life Technologies) genome analyser using the Ion 318 chip and Ion PGM™ Sequencing 400 Kit v2 (Life Technologies) according to the manufacturer's instructions. To separate non-mate reads and split true mate reads, we used the matePairingSplitReads.py script.

Illumina mate-pair sequencing. A mate-pair library with 8-12 kb fragments was prepared from the pure genomic DNA using the Nextera Mate Pair Library Sample Prep Kit (Illumina) and TruSeq DNA Sample Prep Kit (Illumina) according to the manufacturer's recommendations. The library was sequenced on the MiSeq platform (Illumina) using a 2 × 150 cycle MiSeq V2 Reagent Kit according to the standard Illumina sequencing protocols. Demultiplexing was performed using bcl2fastq v2.17.1.14 Conversion Software (Illumina). Adaptor sequences were removed from reads during demultiplexing. For trimming and separation of the single-end, paired-end, and mate-pair reads, NxTrim software was used.

Read datasets corresponding to the three shotgun libraries were combined, and SPAdes 3.6.0 software [17] was used to create a single genome assembly. For eukaryotic contig scaffolding, we used Sspace software [19] with the parameters -p 1 -x 0 -l library.txt -s Contigs.fasta -k 2.

Contigs binning. *Assembly analysis.* The k-mer coverage and contig length were taken from the SPAdes assembly information (contig names). GC content was calculated using the infoseq tool built into the EMBOSS package. Contigs with a length less than 500 bp were ignored. The tetranucleotide content was calculated using calc.kmerfreq.pl, and the script is available at [https://github.com/MadsAlbertsen/miscperlscripts/blob/master/calc.kmerfreq.pl].

The reads of the individual shotgun libraries and the *H. medicinalis* cDNA reads (see below) were mapped against the genome assembly by bowtie2 [91]. The depth of read coverage for the individual libraries was calculated using BEDTools [92]. The taxonomic classification of the contigs was carried out by MEGAN6 [93] using the results of BLAST analysis (nr/nt database).

Classification of eukaryotic contigs. Using the R language and the car package, a concentration ellipse was built to confine at least 99% of those contigs against which at least ten cDNA reads had been mapped. To avoid the loss of potential eukaryotic contigs, we also considered the taxonomic affiliation of those contigs against which cDNA reads were not mapped. All contigs that were identified as neither prokaryotic nor eukaryotic but belonged to the ellipse, were assigned to be eukaryotic.

Annotation of a draft genome. To annotate a draft genome, three sets of so-called hints, sequences in the genome that exhibit features of specific gene structures, such as exons, introns, etc., were generated (for details, see **Supplementary Methods**). The first set of hints was generated using sequences from the *H. robusta* protein coding genes. The second set of hints was generated using contigs corresponding to the *de novo* transcriptome assembly (see below). The third set of hints was generated using the cDNA reads (see below). All sets of hints were combined, and AUGUSTUS software [94] (version 3.7.1) was used for annotation of the draft genome.

Laser microdissection. Laser microdissection was applied both to obtain tissue-specific mRNA samples from salivary cells and muscles for subsequent differential expression analysis and to collect different parts of the digestive tract (crop, coeca, intestine) for consecutive metagenomic analysis (for details, see **Supplementary Methods**). Briefly, live leeches were snap frozen in liquid nitrogen. The different parts of the leech bodies were cryosectioned, and slices were attached to membrane slides with the PEN foil (Leica Microsystems, Germany). Slides were stained with methylene blue to reveal salivary cell bodies. Staining was omitted when the parts of the digestive tract were isolated. Tissue collection was performed for three leech species, *H. verbana*, *H. orientalis* and *H. medicinalis,* by a Leica Laser Microdissection System LMD7000 (Leica Microsystems, Germany). Salivary cells and muscles were isolated directly into RNeasy extraction solution for the subsequent extraction of total RNA, and the parts of the digestive tract were isolated directly into ATL buffer from the QIAamp DNA Micro Kit for the subsequent extraction of total DNA.

cDNA library construction and sequencing. Total RNA was extracted from the tissue fragments isolated by laser microdissection using an ExtractRNA Kit (Evrogen, Russia) (for details, see **Supplementary Methods**). Tissue-specific cDNA libraries were created using the Mint-2 cDNA Synthesis Kit (Evrogen, Russia) according to the manufacturer's instructions. The adapters PlugOligo-3M and CDS-4M were used for cDNA synthesis. The normalization of cDNA was performed using the DNS (duplex-specific nuclease) in accordance with the manufacturer's protocol (Evrogen, Russia). Normalized and non-normalized cDNA libraries (100 ng of each sample) were fragmented to a mean size of 400-500 bp by using a Covaris S220 System (Covaris, Woburn, Massachusetts, USA). Then, an Ion Xpress Plus Fragment Library Kit (Life Technologies) was employed to prepare barcoded shotgun libraries. To conduct emulsion PCR, an Ion PGM Template OT2 400 Kit (Life Technologies) was utilized. Sequencing was performed by the Ion Torrent PGM (Life Technologies) analyser using Ion

318 chips and Ion PGM sequencing 400 Kit v2 (Life Technologies) in accordance with the manufacturer's protocol.

Transcriptome assembly and annotation. Cutadapt [95] v1.9 was applied to trim the adapter sequences used for cDNA synthesis. Prinseq lite [96] v.0.20.4 was used to trim reads according to their quality and length. For *de novo* transcriptome assembly, we used Trinity software [20] (version r20131110) with the default parameters. Blast2GO [21] was used for Gene Ontology (GO) analysis and the functional annotation of contigs. Local BLAST (BlastX threshold value of e = $1x10^{-6}$, matrix BLOSUM-62) and the nr database were used for the grouping and annotation of contigs. MEGAN6 software [93] was used to visualize the contig distribution (by KOG/EGGNOG classifications). TransDecoder and Trinotate were applied to identify and annotate ORFs.

Differential expression analysis. Differentially expressed genes were detected according to a recent protocol using the Bowtie [91], Htseq [97] and edgeR [98] software packages. Because biological replicates were available only for *H. verbana*, we applied the dispersion estimates for *H. verbana* to other leech species. Genes with q-value (FDR) < 0.05 were defined as differentially expressed. To perform differential expression analysis using a genome model, the cDNA reads of *H. medicinalis* were mapped against the genome assembly using the STAR software [99]. In addition to the cDNA reads corresponding to the salivary cells and muscles, we also mapped reads of *H. medicinalis* ganglion 2 against its genome assembly (SRR799260, SRR799263, SRR799266). The HTSeq software [97] was used to count the number of reads mapped against the annotated genes. For each tissue (salivary cells, muscles, and neural tissue), we established a list in which the numbers of mapped reads corresponded to individual gene IDs by applying the "htseq-count" python script with the default parameters. Unique transcripts corresponding to certain tissues were determined by finding the intersection of these lists.

Collection and concentration of the salivary cell secretions. To collect SCSs from three medicinal leech species, the bottom of a 15 mL Falcon tube was cut off, and an impermeable membrane (PARAFILM® M) was stretched on the excised end. The tube was filled with saline solution containing 10 mM arginine. A leech bit through the membrane, sucked up the salt solution and emitted its secretion into the saline solution. The saline solution enriched with SCS was continuously stirred and renewed to prevent its ingestion by leech. We collected approximately 10 mL of saline solution containing highly diluted leech SCS. Harvested SCSs were concentrated on a solid-phase extraction Sep-Pak Vac C18 6cc cartridge (Waters, USA) using 0.1% TFA in 70%/30% acetonitrile/water (v/v) as the buffer for elution of the protein fraction. The acetonitrile was evaporated, and the remaining solution was lyophilized. Dry protein powder was stored at -70°C prior to mass spectrometric analysis.

Digestion of the salivary cell secretions and sample preparation for proteomic analysis. Several digestion and preparation methods were applied to each freeze-dried sample of the SCS to cover the broadest possible variety of the salivary proteins. These preparation methods included filter-aided sample preparation (FASP), gel-free trypsin digestion using surfactant RapiGest SF (Waters), and in-gel digestion of the protein sample. Detailed descriptions of the preparation methods are presented in the **Supplementary Methods.**

Mass spectrometry and analysis of mass spectra. Mass spectra were obtained for each prepared protein sample by using three instruments: (i) a TripleTOF 5600+ mass spectrometer with a NanoSpray III ion source (Sciex, USA) coupled to a NanoLC Ultra 2D+ nano-HPLC system (Eksigent, USA), (ii) a Q-Exactive HF mass spectrometer with a nanospray Flex ion source (Thermo Scientific,

Germany), and (iii) a Maxis 3G mass spectrometer with an HDC-cell upgrade and an Online NanoElectrospray ion source (Bruker Daltonics GmbH, Germany) coupled to a Dionex Ultimate 3000 (ThermoScientific, USA) HPLC system. The obtained raw mass spectra were converted into non-calibrated peaklists by the appropriate software, and these peaklists were analysed using ProteinPilot 4.5 revision 1656 (ABSciex). The acquisition of the mass spectra and their analysis and peptide identification are described in detail in the **Supplementary Methods**.

To create a final list of the identified proteins for each medicinal leech species, the combined transcriptomes were translated either by ORF Transdecoder (standard parameters) or by the six-reading-frame method (length filter was ≥30 aa). Then, the protein sequences obtained by both methods were combined and used to establish a referential database of potential SCS proteins for each medicinal leech species. The protein sequences in the referential database that were matched by the peptides identified in an individual peaklist allowed the creation of protein datasets for individual samples. The protein datasets generated for individual samples by using different preparation methods and mass spectrometry techniques were combined to create the final list of the SCS proteins for each medicinal leech species.

Data Access

All of the raw reads generated in this study have been deposited in the NCBI database under BioProject accessions PRJNA257563 and PRJNA256119.

Acknowledgments

H. medicinalis genome sequence, assembly and annotation were supported by the Russian Science Foundation (project №17-75-20099).

Transcriptomic analysis of salivary and muscle cells of *H. medicinalis*, *H. verbana*, *H. orientalis* and proteomic analysis of their salivary cell secretions were supported by the Russian Science Foundation (project №14-14-00696).

O.V. Podgorny was partially supported by the IDB RAS Government basic research program, no. 0108-2018-0007. Laser microdissection was performed using equipment of the Core Centrum of Institute of Developmental Biology RAS.

V.G.Z mass spectrometric measurements were performed using the equipment of “Human Proteome” Core Facility of the Orekhovich Institute of Biomedical Chemistry (Russia) which is supported by Ministry of Education and Science of the Russian Federation (agreement 14.621.21.0017, unique project ID RFMEFI62117X0017)

M.S.G. is grateful to the Russian Science Foundation (grant No. 14-50-00150) for support.

We thank, Alexander V. Bazin the director of the company HIRUD I.N. Ltd.

Author contributions

V.N.L., O.V. Podgorny and V.V.B. conceived and designed the experiments. I.P.B., V.A.M., A.S. Kurdyumov, and D.A.S. samples collection. O.V. Podgorny, D.A.S. and V.V.B. laser microdissection.
V.V.B., D.A.S., and V.N.L. generated DNA and RNA for sequencing. V.V.B., M.T.V. and E.S.K. genome and transcriptome sequencing. V.V.B., A.S. Kasianov, and A.I.M. genome assembly, binning and automated annotation. V.V.B., A.S. Kasianov, and D.V.V. transcriptome assembly and annotation. D.V.V., A.S.N. differential expression analysis. V.G.Z., S.I.K., N.A.A., I.O.B., O.V. Pobeguts, D.S.M and D.V.R. sample preparation and mass spectrometry for proteomic analysis. V.V.B., E.N.G. analyzed results and made the figures. V.V.B. and O.V. Podgorny wrote the paper. I.P.B., M.S.G., V.M.G., H.B.S. and V.N.L. coordinated and revised the paper.

Conflicts of Interest

The authors declare no competing financial interests.

Corresponding author

Correspondence to Vladislav V. Babenko (email address daniorerio34@gmail.com)

Reference

1. Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, et al. Genome sequence of the Asian Tiger mosquito, Aedes albopictus , reveals insights into its biology, genetics, and evolution. Proc Natl Acad Sci. 2015;112:E5907–15. doi:10.1073/pnas.1516410112.

2. Mesquita RD, Vionette-Amaral RJ, Lowenberger C, Rivera-Pomar R, Monteiro FA, Minx P, et al. Genome of Rhodnius prolixus , an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. Proc Natl Acad Sci. 2015;112:14936–41. doi:10.1073/pnas.1506226112.

3. Zepeda Mendoza ML, Xiong Z, Escalera-Zamudio M, Runge AK, Thézé J, Streicker D, et al. Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. Nat Ecol Evol. 2018;2:659–68. doi:10.1038/s41559-018-0476-8.

4. Gulia-Nuss M, Nuss AB, Meyer JM, Sonenshine DE, Roe RM, Waterhouse RM, et al. Genomic insights into the Ixodes scapularis tick vector of Lyme disease. Nat Commun. 2016;7:10507. doi:10.1038/ncomms10507.

5. Rosenfeld JA, Reeves D, Brugler MR, Narechania A, Simon S, Durrett R, et al. Genome assembly and geospatial phylogenomics of the bed bug Cimex lectularius. Nat Commun. 2016;7:10164. doi:10.1038/ncomms10164.

6. Hyson JM. Leech therapy: a history. J Hist Dent. 2005;53:25–7.

7. Houschyar KS, Momeni A, Maan ZN, Pyles MN, Jew OS, Strathe M, et al. Medical leech therapy in plastic reconstructive surgery. Wiener Medizinische Wochenschrift. 2015;165:419–25. doi:10.1007/s10354-015-0382-5.

8. Basanova A V, Baskova IP, Zavalova LL. Vascular-platelet and plasma hemostasis regulators from bloodsucking animals. Biochemistry (Mosc). 2002;67:143–50.

9. Hildebrandt J-P, Lemke S. Small bite, large impact–saliva and salivary molecules in the medicinal leech, Hirudo medicinalis. Naturwissenschaften. 2011;98:995–1008. doi:10.1007/s00114-011-0859-z.

10. Rigbi M, Orevi M, Eldor A. Platelet aggregation and coagulation inhibitors in leech saliva and their roles in leech therapy. Semin Thromb Hemost. 1996;22:273–8.

11. Lemke S, Müller C, Lipke E, Uhl G, Hildebrandt J-P. May Salivary Gland Secretory Proteins from Hematophagous Leeches (Hirudo verbana) Reach Pharmacologically Relevant Concentrations in the Vertebrate Host? PLoS One. 2013;8:e73809. doi:10.1371/journal.pone.0073809.

12. Lemke S, Müller C, Hildebrandt J-P. Be ready at any time: postprandial synthesis of salivary proteins in salivary gland cells of the haematophagous leech Hirudo verbana. J Exp Biol. 2016;219 Pt 8:1139–45. doi:10.1242/jeb.135509.

13. Siddall ME, Brugler MR, Kvist S. Comparative Transcriptomic Analyses of Three Species of Placobdella (Rhynchobdellida: Glossiphoniidae) Confirms a Single Origin of Blood Feeding in Leeches. J Parasitol. 2016;102:143–50. doi:10.1645/15-802.

14. Kvist S, Min G-S, Siddall ME. Diversity and selective pressures of anticoagulants in three medicinal leeches (Hirudinida: Hirudinidae, Macrobdellidae). Ecol Evol. 2013;3:918. doi:10.1002/ECE3.480.

15. Kvist S, Brugler MR, Goh TG, Giribet G, Siddall ME. Pyrosequencing the salivary transcriptome of Haemadipsa interrupta (Annelida: Clitellata: Haemadipsidae): anticoagulant diversity and insight into the evolution of anticoagulation capabilities in leeches. Invertebr Biol. 2014;133:74–98. doi:10.1111/ivb.12039.

16. Amorim AMXP, De Oliveira UC, Faria F, Pasqualoto KFM, Junqueira-De-Azevedo IDLM, Chudzinski-Tavassi AM. Transcripts involved in hemostasis: Exploring salivary complexes from Haementeria vizottoi leeches through transcriptomics, phylogenetic studies and structural features. Toxicon. 2015;106:20–9.

17. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19:455–77. doi:10.1089/cmb.2012.0021.

18. Nikitina A, Babenko V, Akopian T, Shirokov D, Manuvera V, Kurdyumov A, et al. Draft mitochondrial genomes of Hirudo medicinalis and Hirudo verbana (Annelida, Hirudinea). Mitochondrial DNA Part B Resour. 2016;1:254–6. doi:10.1080/23802359.2016.1157774.

19. Boetzer M, Henkel C V., Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27:578–9. doi:10.1093/bioinformatics/btq683.

20. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52. doi:10.1038/nbt.1883.

21. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a

universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6. doi:10.1093/bioinformatics/bti610.

22. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 2016;44:D286–93. doi:10.1093/nar/gkv1248.

23. Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, et al. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc. 2013;8:1765–86. doi:10.1038/nprot.2013.099.

24. Hibsh D, Schori H, Efroni S, Shefi O. De novo transcriptome assembly databases for the central nervous system of the medicinal leech. Sci Data. 2015;2:150015. doi:10.1038/sdata.2015.15.

25. Rigbi M, Levy H, Iraqi F, Teitelbaum M, Orevi M, Alajoutsijärvi A, et al. The saliva of the medicinal leech Hirudo medicinalis--I. Biochemical characterization of the high molecular weight fraction. Comp Biochem Physiol B. 1987;87:567–73.

26. Baskova IP, Yudina TG, Zavalova LL, Dudkina AS. Protein-lipid particles of medicinal leech salivary gland secretion; their size and morphology. Biochemistry (Mosc). 2010;75:585–9.

27. Seals DF, Courtneidge SA. The ADAMs family of metalloproteases: multidomain proteins with multiple functions. Genes Dev. 2003;17:7–30. doi:10.1101/gad.1039703.

28. Kini R, Koh C. Metalloproteases Affecting Blood Coagulation, Fibrinolysis and Platelet Aggregation from Snake Venoms: Definition and Nomenclature of Interaction Sites. Toxins (Basel). 2016;8:284. doi:10.3390/toxins8100284.

29. Chagas AC, Calvo E, Rios-Velásquez CM, Pessoa FA, Medeiros JF, Ribeiro JM. A deep insight into the sialotranscriptome of the mosquito, Psorophora albipes. BMC Genomics. 2013;14:875. doi:10.1186/1471-2164-14-875.

30. Ribeiro JMC, Slovák M, Francischetti IMB. An insight into the sialome of Hyalomma excavatum. Ticks Tick Borne Dis. 2017;8:201–7. doi:10.1016/j.ttbdis.2016.08.011.

31. Kini RM. Anticoagulant proteins from snake venoms: structure, function and mechanism. Biochem J. 2006;397:377–87. doi:10.1042/BJ20060302.

32. Brahma RK, McCleary RJR, Kini RM, Doley R. Venom gland transcriptomics for identifying, cataloging, and characterizing venom proteins in snakes. Toxicon. 2015;93:1–10. doi:10.1016/j.toxicon.2014.10.022.

33. Bland ND, Pinney JW, Thomas JE, Turner AJ, Isaac RE. Bioinformatic analysis of the neprilysin (M13) family of peptidases reveals complex evolutionary and functional relationships. BMC Evol Biol. 2008;8:16. doi:10.1186/1471-2148-8-16.

34. Modica MV, Lombardo F, Franchini P, Oliverio M. The venomous cocktail of the vampire snail Colubraria reticulata (Mollusca, Gastropoda). BMC Genomics. 2015;:1–21. doi:10.1186/s12864-015-1648-4.

35. Rivière G, Michaud A, Deloffre L, Vandenbulcke F, Levoye A, Breton C, et al. Characterization of the first non-insect invertebrate functional angiotensin-converting enzyme (ACE): leech TtACE resembles the N-domain of mammalian ACE. Biochem J. 2004;382 Pt 2:565–73. doi:10.1042/BJ20040522.

36. Gingras R, Richard C, El-Alfy M, Morales CR, Potier M, Pshezhetsky A V. Purification, cDNA cloning, and expression of a new human blood plasma glutamate carboxypeptidase homologous to N-acetyl-aspartyl-alpha-glutamate carboxypeptidase/prostate-specific membrane antigen. J Biol Chem. 1999;274:11742–50.

37. Sojka D, Franta Z, Horn M, Caffrey CR, Mareš M, Kopáček P. New insights into the machinery of blood digestion by ticks. Trends Parasitol. 2013;29:276–85. doi:10.1016/j.pt.2013.04.002.

38. Reverter D, Vendrell J, Canals F, Horstmann J, Avilés FX, Fritz H, et al. A carboxypeptidase inhibitor from the medical leech Hirudo medicinalis. Isolation, sequence analysis, cDNA cloning, recombinant expression, and characterization. J Biol Chem. 1998;273:32927–33.

39. Nielsen VG, Crow JP, Mogal A, Zhou F, Parks DA. Peroxynitrite decreases hemostasis in human plasma in vitro. Anesth Analg. 2004;99:21–6.

40. Kurahashi T, Fujii J. Roles of Antioxidative Enzymes in Wound Healing. J Dev Biol 2015, Vol 3, Pages 57-70. 2015;3:57–70. doi:10.3390/JDB3020057.

41. Crispell G, Budachetri K, Karim S. Rickettsia parkeri colonization in Amblyomma maculatum: The role of superoxide dismutases. Parasites and Vectors. 2016;9:1–12. doi:10.1186/s13071-016-1579-1.

42. Graça-Souza A V, Maya-Monteiro C, Paiva-Silva GO, Braz GRC, Paes MC, Sorgine MHF, et al. Adaptations against heme toxicity in blood-feeding arthropods. Insect Biochem Mol Biol. 2006;36:322–35. doi:10.1016/j.ibmb.2006.01.009.

43. Tripp BC, Smith K, Ferry JG. Carbonic anhydrase: new insights for an ancient enzyme. J Biol Chem. 2001;276:48615–8. doi:10.1074/jbc.R100045200.

44. Carter MJ, Parsons DS. The isoenzymes of carbonic anhydrase: tissue, subcellular distribution and functional significance, with particular reference to the intestinal tract. J Physiol. 1971;215:71–94.

45. Santos VC, Nunes CA, Pereira MH, Gontijo NF. Mechanisms of pH control in the midgut of Lutzomyia longipalpis: roles for ingested molecules and hormones. J Exp Biol. 2011;214 Pt 9:1411–8. doi:10.1242/jeb.051490.

46. Linser PJ, Smith KE, Seron TJ, Neira Oviedo M. Carbonic anhydrases and anion transport in mosquito midgut pH regulation. J Exp Biol. 2009;212 Pt 11:1662–71. doi:10.1242/jeb.028084.

47. Hovingh P, Linker A. Hyaluronidase activity in leeches (Hirudinea). Comp Biochem Physiol B Biochem Mol Biol. 1999;124:319–26.

48. Kemparaju K, Girish KS. Snake venom hyaluronidase: a therapeutic target. Cell Biochem Funct. 2006;24:7–12. doi:10.1002/cbf.1261.

49. Shriver Z, Sundaram M, Venkataraman G, Fareed J, Linhardt R, Biemann K, et al. Cleavage of the antithrombin III binding site in heparin by heparinases and its implication in the generation of low molecular weight heparin. Proc Natl Acad Sci U S A. 2000;97:10365–70.

50. Yip J, Shen Y, Berndt MC, Andrews RK. Primary platelet adhesion receptors. IUBMB Life. 2005;57:103–8.

51. Ribeiro JM, Charlab R, Valenzuela JG. The salivary adenosine deaminase activity of the mosquitoes Culex quinquefasciatus and Aedes aegypti. J Exp Biol. 2001;204 Pt 11:2001–10.

52. Modica MV, Lombardo F, Franchini P, Oliverio M. The venomous cocktail of the vampire snail Colubraria reticulata (Mollusca, Gastropoda). BMC Genomics. 2015;16:441. doi:10.1186/s12864-015-1648-4.

53. Burnstock G, Wood JN. Purinergic receptors: Their role in nociception and primary afferent neurotransmission. Curr Opin Neurobiol. 1996;6:526–32.

54. Seymour JL, Henzel WJ, Nevins B, Stults JT, Lazarus RA. Decorsin. A potent glycoprotein IIb-IIIa antagonist and platelet aggregation inhibitor from the leech Macrobdella decora. J Biol Chem. 1990;265:10143–7.

55. Pang SS, Wijeyewickrema LC, Hor L, Tan S, Lameignere E, Conway EM, et al. The Structural Basis for Complement Inhibition by Gigastasin, a Protease Inhibitor from the Giant Amazon Leech. J Immunol. 2017;:ji1700158. doi:10.4049/jimmunol.1700158.

56. Yamazaki Y, Morita T. Structure and function of snake venom cysteine-rich secretory proteins. Toxicon. 2004;44:227–31. doi:10.1016/J.TOXICON.2004.05.023.

57. Koludarov I, Jackson TNW, Sunagar K, Nouwens A, Hendrikx I, Fry BG. Fossilized venom: the unusually conserved venom profiles of Heloderma species (beaded lizards and gila monsters). Toxins (Basel). 2014;6:3582–95. doi:10.3390/toxins6123582.

58. Chhabra S, Chang SC, Nguyen HM, Huq R, Tanner MR, Londono LM, et al. Kv1.3 channel-blocking immunomodulatory peptides from parasitic worms: Implications for autoimmune diseases. FASEB J. 2014;28:3952–64.

59. Ma D, Francischetti IMB, Ribeiro JMC, Andersen JF. The structure of hookworm platelet inhibitor (HPI), a CAP superfamily member from Ancylostoma caninum. Acta Crystallogr Sect F, Struct Biol Commun. 2015;71 Pt 6:643–9. doi:10.1107/S2053230X1500271X.

60. Shikamoto Y, Suto K, Yamazaki Y, Morita T, Mizuno H. Crystal structure of a CRISP family Ca2+ -channel blocker derived from snake venom. J Mol Biol. 2005;350:735–43. doi:10.1016/j.jmb.2005.05.020.

61. Wang Y-L, Kuo J-H, Lee S-C, Liu J-S, Hsieh Y-C, Shih Y-T, et al. Cobra CRISP functions as an inflammatory modulator via a novel Zn2+- and heparan sulfate-dependent transcriptional regulation of endothelial cell adhesion molecules. J Biol Chem. 2010;285:37872–83. doi:10.1074/jbc.M110.146290.

62. Rawlings ND, Tolle DP, Barrett AJ. Evolutionary families of peptidase inhibitors. Biochem J. 2004;378 Pt 3:705–16. doi:10.1042/BJ20031825.

63. Grzonka Z, Jankowska E, Kasprzykowski F, Kasprzykowska R, Lankiewicz L, Wiczk W, et al. Structural studies of cysteine proteases and their inhibitors. Acta Biochim Pol. 2001;48:1–20.

64. Wang Y, Zhou Y, Gong H, Cao J, Zhang H, Li X, et al. Functional characterization of a cystatin from the tick Rhipicephalus haemaphysaloides. Parasit Vectors. 2015;8:140. doi:10.1186/s13071-015-0725-5.

65. Chmelař J, Kotál J, Langhansová H, Kotsyfakis M. Protease Inhibitors in Tick Saliva: The Role of Serpins and Cystatins in Tick-host-Pathogen Interaction. Front Cell Infect Microbiol. 2017;7:216. doi:10.3389/fcimb.2017.00216.

66. Dainichi T, Maekawa Y, Ishii K, Zhang T, Nashed BF, Sakai T, et al. Nippocystatin, a cysteine protease inhibitor from Nippostrongylus brasiliensis, inhibits antigen processing and modulates antigen-specific immune response. Infect Immun. 2001;69:7380–6. doi:10.1128/IAI.69.12.7380-7386.2001.

67. Ho DH, Badellino K, Baglia FA, Walsh PN. A binding site for heparin in the apple 3 domain of factor XI. J Biol Chem. 1998;273:16382–90. doi:10.1074/JBC.273.26.16382.

68. Brown PJ, Gill AC, Nugent PG, McVey JH, Tomley FM. Domains of invasion organelle proteins from apicomplexan parasites are homologous with the Apple domains of blood coagulation factor XI and plasma pre-kallikrein and are members of the PAN module superfamily. FEBS Lett. 2001;497:31–8.

69. Huizinga EG, Schouten A, Connolly TM, Kroon J, Sixma JJ, Gros P. The structure of leech anti-platelet protein, an inhibitor of haemostasis. Acta Crystallogr D Biol Crystallogr. 2001;57 Pt 8:1071–8.

70. Rehman AA, Ahsan H, Khan FH. α-2-Macroglobulin: a physiological guardian. J Cell Physiol. 2013;228:1665–75. doi:10.1002/jcp.24266.

71. Cvirn G, Gallistl S, Muntean W. Effects of alpha(2)-macroglobulin and antithrombin on thrombin generation and inhibition in cord and adult plasma. Thromb Res. 2001;101:183–91.

72. Fujita T. Evolution of the lectin–complement pathway and its role in innate immunity. Nat Rev Immunol. 2002;2:346–53. doi:10.1038/nri800.

73. Angthong P, Roytrakul S, Jarayabhand P, Jiravanichpaisal P. Characterization and function of a tachylectin 5-like immune molecule in Penaeus monodon. Dev Comp Immunol. 2017;76:120–31. doi:10.1016/j.dci.2017.05.023.

74. Kairies N, Beisel H-G, Fuentes-Prior P, Tsuda R, Muta T, Iwanaga S, et al. The 2.0-A crystal structure of tachylectin 5A provides evidence for the common origin of the innate immunity and the blood coagulation systems. Proc Natl Acad Sci. 2001;98:13519–24. doi:10.1073/pnas.201523798.

75. OmPraba G, Chapeaurouge A, Doley R, Devi KR, Padmanaban P, Venkatraman C, et al. Identification of a Novel Family of Snake Venom Proteins Veficolins from *Cerberus rynchops* Using a Venom Gland Transcriptomics and Proteomics Approach. J Proteome Res. 2010;9:1882–93. doi:10.1021/pr901044x.

76. Kane WH, Davie EW. Blood coagulation factors V and VIII: structural and functional

similarities and their relationship to hemorrhagic and thrombotic disorders. Blood. 1988;71:539–55.

77. Foster PA, Fulcher CA, Houghten RA, Zimmerman TS. Synthetic factor VIII peptides with amino acid sequences contained within the C2 domain of factor VIII inhibit factor VIII binding to phosphatidylserine. Blood. 1990;75:1999–2004.

78. Leitinger B. Discoidin Domain Receptor Functions in Physiological and Pathological Conditions. In: International review of cell and molecular biology. 2014. p. 39–87. doi:10.1016/B978-0-12-800180-6.00002-5.

79. Foster PA, Fulcher CA, Houghten RA, Zimmerman TS. Synthetic factor VIII peptides with amino acid sequences contained within the C2 domain of factor VIII inhibit factor VIII binding to phosphatidylserine. Blood. 1990;75:1999–2004.

80. Jeon H, Blacklow SC. STRUCTURE AND PHYSIOLOGIC FUNCTION OF THE LOW-DENSITY LIPOPROTEIN RECEPTOR. Annu Rev Biochem. 2005;74:535–62. doi:10.1146/annurev.biochem.74.082803.133354.

81. Zipser B, Bradford JJ, Hollingsworth RI. Cholesterol and its derivatives, are the principal steroids isolated from the leech species Hirudo medicinalis. Comp Biochem Physiol C Pharmacol Toxicol Endocrinol. 1998;120:269–82.

82. Gupta RK, Gupta GS. R-Type Lectin Families. In: Animal Lectins: Form, Function and Clinical Applications. Vienna: Springer Vienna; 2012. p. 313–30. doi:10.1007/978-3-7091-1065-2_14.

83. Uchida T, Yamasaki T, Eto S, Sugawara H, Kurisu G, Nakagawa A, et al. Crystal Structure of the Hemolytic Lectin CEL-III Isolated from the Marine Invertebrate Cucumaria echinata. J Biol Chem. 2004;279:37133–41. doi:10.1074/jbc.M404065200.

84. Min G-S, Sarkar IN, Siddall ME. Salivary Transcriptome of the North American Medicinal Leech, Macrobdella decora. J Parasitol. 2010;96:1211–21. doi:10.1645/GE-2496.1.

85. Suzuki R, Kuno A, Hasegawa T, Hirabayashi J, Kasai KI, Momma M, et al. Sugar-complex structures of the C-half domain of the galactose-binding lectin EW29 from the earthworm Lumbricus terrestris. Acta Crystallogr Sect D Biol Crystallogr. 2009;65:49–57.

86. Perkins SJ, Smith KF, Williams SC, Haris PI, Chapman D, Sim RB. The Secondary Structure of the von Willebrand Factor type A Domain in Factor B of Human Complement by Fourier Transform Infrared Spectroscopy: Its Occurrence in Collagen Types VI, VII, XII and XIV, the Integrins and Other Proteins by Averaged Structure Pr. J Mol Biol. 1994;238:104–19. doi:10.1006/JMBI.1994.1271.

87. Malhotra A. Mutation, Duplication, and More in the Evolution of Venomous Animals and Their Toxins. Springer, Dordrecht; 2017. p. 33–45. doi:10.1007/978-94-007-6458-3_5.

88. Lukas P, Wolf R, Rauch BH, Hildebrandt J-P, Müller C. Hirudins of the Asian medicinal leech, Hirudinaria manillensis: same same, but different. Parasitol Res. 2019;118:2223–33. doi:10.1007/s00436-019-06365-z.

89. Müller C, Haase M, Lemke S, Hildebrandt J-P. Hirudins and hirudin-like factors in Hirudinidae: implications for function and phylogenetic relationships. Parasitol Res.

2017;116:313–25. doi:10.1007/s00436-016-5294-9.

90. Walz B, Schäffner K-H, Sawyer RT. Ultrastructure of the anterior salivary gland cells of the giant leech, Haementeria ghilianii (Annelida, Hirudinea). J Morphol. 1988;196:321–32. doi:10.1002/jmor.1051960305.

91. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. doi:10.1038/nmeth.1923.

92. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. doi:10.1093/bioinformatics/btq033.

93. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17:377–86. doi:10.1101/gr.5969107.

94. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33 Web Server issue:W465-7. doi:10.1093/nar/gki458.

95. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17:10. doi:10.14806/ej.17.1.200.

96. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27:863–4. doi:10.1093/bioinformatics/btr026.

97. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31:166–9. doi:10.1093/bioinformatics/btu638.

98. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40. doi:10.1093/bioinformatics/btp616.

99. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. doi:10.1093/bioinformatics/bts635.