Far beyond common leeching: insights into an ancient medical device through integrated omics data

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], Sergey I. Kovalchuk[1,7], Nickolay A. Anikanov[1,7], Ivan O. Butenko[1], Maria T. Vakhitova[1,3], Olga V. Pobeguts[1], Daria S. Matushkina[1], Daria V. Rakitina[1], Daria D. Kharlampieva[1], Nadezhda F. Polina[1], Pavel A. Bobrovsky[1], Elena S. Kostryukova[1], Victor G. Zgoda[8], Isolda P. Baskova[9], Mikhail S. Gelfand[5,6,10,11], Vadim M. Govorun[1,3,], 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]Faculty of Computer Science, National Research University Higher School of Economics, 20 Myasnitskaya str., Moscow 101000, Russia

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

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. To fill the gap in our knowledge about the genetically encoded components of leech SCS, 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. In addition, a metagenomic analysis performed in our study revealed the overall metabolic potential of the leech microbiota. Thus, our study extends the knowledge of the genetic fundamentals of the blood-sucking lifestyle in leeches.

Introduction

Medicinal leeches belong to a relatively small group of highly specialized annelids (*Hirudinea*). The group includes approximately 680 species discriminated by morphological characteristics[1]. Most of them (approximately 480 species) are freshwater, and the rest reside in terrestrial and marine environments. Many leeches are blood-sucking ectoparasites, while others are predators. The medicinal leech is a common model organism for neurophysiological studies[2] and is approved by the FDA as a medical device[3].

Blood-sucking leeches have been used for bloodletting to treat a variety of ailments since ancient times[4]. 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, medicinal leeches are used to provide tissue drainage after replantation when the surgical correction of venous congestion is not feasible or fails[5]. 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[6,7]. The composition of leech saliva has been shown to play a key role in this inhibition[7,8].

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 use their jaws to make an incision in the host skin, and the salivary cell secretion (SCS) is released into the wound at the feeding site[7]. In addition to its inhibitory effect on haemostasis and blood coagulation, SCS suppresses inflammation, exhibits analgesic effects, possesses antimicrobial activity, and enhances local blood circulation, facilitating leech feeding. Moreover, some SCS components are thought to preserve the blood from rapid degradation after ingestion.

Although some SCS components have been characterized and their respective targets in hosts have been identified[7], the exact molecular mechanisms of the orchestrated interaction between the SCS and the components of haemostasis are still poorly understood. Therefore, elucidating the genetic basis of the SCS composition will illuminate the molecular mechanisms that underlie the beneficial effects of hirudotherapy and facilitate the development of novel pharmacological compounds to treat impaired peripheral blood circulation, venous congestion or microbial infections.

Medicinal leeches, which live exclusively on blood, must solve two main problems associated with blood meals. The first is a lack of certain nutrients necessary for organism functioning, and the second is the release of toxic compounds, particularly haem, during digestion of the blood. The bacterial community in the leech digestive system is essential for solving these problems. Although many studies have described the diversity of the leech microbiome, little is known about the functional contributions of bacterial proteins to digestion, absorption and host maintenance.

In the current study, we performed transcriptional profiling of the salivary cells followed by proteomic validation to fill a gap in our knowledge about the genetically encoded components of the SCSs of three medicinal leeches, *Hirudo medicinalis, Hirudo orientalis*, and *Hirudo verbana*. We performed *de novo* assembly of the *H. medicinalis* genome and used it to describe the blood meal-related genes. Finally, we compared the taxonomic composition of the crop and intestine microbiota and performed function-related metagenomic analysis to identify bacterial genes involved in digestion, absorption and host maintenance.

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. Ion Proton sequencing produced a read dataset with a mean read length of 200 bp. After the reads were filtered by quality (≥Q20) and length (≥200 bp), the total read length was 11.3 Gbp, which corresponds to a theoretical coverage of ×50. Mate-pair sequencing by Ion Torrent and Illumina produced two read datasets with ×2.5 and ×10 coverage, respectively (**Supplementary Table 1**). All read datasets were combined, and a single assembly was created[9]. 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 combination of the BlastN algorithm and the 'nt' database allowed the use of the lowest common ancestor (LCA) method (MEGAN6 software) to determine the taxonomic membership of the contigs[10] (**Supplementary Data 1**). We combined two standard approaches to separate and cluster metagenomic sequences[11]. The discrimination of the eukaryotic and prokaryotic contigs is illustrated in **Supplementary Table, Fig. 1A** and **Supplementary Data 2.** Additionally, we selected the mitochondrial contigs to assemble the leech mitochondrial genome[12].

The eukaryotic contigs underwent a scaffolding procedure using paired reads. Scaffolds were generated using Illumina paired-end and mate-pair read datasets by SSPACE[13]. 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 (**Supplementary Figs. 1-7**).

Prior to annotation, the scaffolds were examined for repetitive elements using RepeatMasker and the invertebrate repetition database presented in Repbase-GIRI[14]. Repetitive elements occupy 12% of the leech genome (**Fig. 1B**, **Supplementary Table 7**). Transposable elements (TEs) of the two main classes (Class I and II) occupy 4.8% of the genome in almost equal proportions. Little is known about TEs in the genomes of annelids[15]; however, these results are consistent with those in other well-known representatives of *Lophotrochozoa*, for example, the annelids *Helobdella robusta* and *Capitella teleta* and the molluscs *Lottia gigantea* and *Octopus bimaculoides*[15,16]. DNA transposons occupy 2.4% of the genome. The most abundant families of DNA transposons are EnSpm, hAT, and Crypton, and the absence of prokaryotic DNA transposons in the leech genome assembly confirms the accuracy of the binning procedure. The leech genome contains approximately 1.3% LTR and 1% non-LTR retrotransposons. Most of the LTR retrotransposons belong to the Gypsy family, while most of the non-LTR retrotransposons belong to the CR1, L2, and Daphne families.

Using OrthoFinder[17], we analysed orthologous gene clusters among representatives of annelids (*H. medicinalis*, *Helobdella robusta*, *Capitella teleta)* and other related taxa in *Lophotrochozoa* (*Brachiopoda, Lingula anatina* and molluscs *O. bimaculoides*, *Aplysia californica*). Two vertebrate species, *Danio rerio* and *Homo sapiens*, served as an outgroup (**Fig. 1C**).

We also identified new homologs of genes encoding known anticoagulants or blood meal-related proteins (**Supplementary Figs. 8-9 Table 8,15**).

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 9**). 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[18] (**Supplementary Table 10**). 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[19] 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 10**). The taxonomy distribution of the closest Blast hits also was similar (**Supplementary Figure 11**). 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[20] (**Supplementary Figure 12**). 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 13**).

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[21]. 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 14**). 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 11-12**). 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[22] (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 13**).

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 2 months. The SCSs were collected according to a previously reported method[8] 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[7] and contains proteinase inhibitors, glycoprotein complexes, and lipids. The latter may form complexes with proteins[23]. 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 14**). Combination of the transcriptomic and proteomic data revealed 25 orthologous clusters of genes unique to the salivary cells (**Supplementary Table 12**). 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 the observations made by the differential expression analysis. 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 (**Supplementary Table 15**).

Leech metagenome

Using binning, we were able to identify bacterial contigs in the *H. medicinalis* genome assembly (**Supplementary Table 3**). Analysis of these contigs yielded taxonomic groups abundant in the medicinal leech metagenome, *Burkholderiales, Rhizobiales, Bacteroidales, Sphingobacteriales, Myxococcales, Chitinophagales,* and *Gammaproteobacteria.* Other identified taxonomic groups were less abundant in the medicinal leech metagenome, and their contigs did not form discrete clusters (**Figs. 4A, B**). The contig size distribution and sequence coverage of the largest bins allowed us to estimate the abundance of the main taxonomic groups in the leech metagenome (**Supplementary Fig. 15**). The analysis of contigs that contained 16S rRNA genes identified 10 full-length ribosomal genes. We performed their taxonomic classification (**Supplementary Tables 16, 17**) and found that two of the identified 16S rRNA genes belong to *Niabella sp.*[24] and *Mucinivorans hirudinis*[25], which have been previously shown to present exclusively in leech metagenomes. Some of the 16S rRNA genes were unknown or belonged to unknown soil bacteria.

Next, we sought to determine the taxonomic composition of the digestive system, as it represents the largest reservoir of bacteria in leeches. To understand the role of bacteria in the digestion, absorption and storage of the blood as well as in host maintenance, we performed taxonomic and function-related metagenomic analysis of distinct compartments of the leech digestive system. Using laser microdissection, we isolated tissue fragments of the crop, coeca, and intestine from three medicinal

leech species, *H. medicinalis, H. orientalis*, and *H. verbana*. For each medicinal leech species, total DNA was extracted from the tissue fragments, and variable region V3-V4 of the 16S rRNA gene was amplified and sequenced by MiSeq (Illumina). We also amplified and sequenced variable region V3-V4 of the 16S rRNA gene from the *H. medicinalis* DNA sample, which had been used to create the shotgun libraries for genome sequencing. Analysis of the read distribution revealed that the microbiome compositions of the crop, coeca, and intestine were similar among the three medicinal leech species (**Fig. 5A, Supplementary Figs. 16, 17**). The crop and coeca had similar bacterial compositions, which differed substantially from the microbial community of the intestine. The coeca is a protrusion of the crop, and the similarity of the bacterial compositions was expected. Therefore, we performed further comparative metagenome analysis of the crop and the intestine. The microbial community of the crop was found to be more diverse than that of the intestine. The crop primarily contained bacterial taxa such as *Burkholderiales, Pseudomonadales, Rhizobiales, Sphingobacteriales*, while the taxonomic groups *Actinomycetales, Enterobacteriales, Bdellovibrionales,* and *Neisseriales* had low representation in the crop.

Reads corresponding to *Myxococcales* members were abundantly represented in the DNA sample of the *H. medicinalis* crop ('crop_med'). This is consistent with the analysis of the contig data of the *H. medicinalis* shotgun library. The intestine microbiota primarily consists of *Bacteroidales* (genus *Mucinivorans*), while *Rhodospirillales* and *Desulfovibrionales* are less abundant. Although the presence of *Aeromonadales* in medicinal leech metagenomes had been previously demonstrated by several research groups[26,27], we could detect their abundant representation in only two samples, the *H. verbana* intestine ('intestine_ver') and the *H. medicinalis* shotgun library ('shotgun').

To build metabolic maps and perform function-related metagenome analysis, we divided the bacterial contigs corresponding to the *H. medicinalis* shotgun library into two groups according to the results of 16S rRNA gene sequencing. One group corresponded to the intestine sample, while the other corresponded to the combined crop and coeca samples. Both contig groups were annotated using PROKKA[28] and GhostKOALA[29] software. Taxonomic analysis showed that contigs of the shotgun library predominantly matched the crop contigs, and only *Mucinivorans hirudinus*, a member of the *Bacteroidales* filum, was identified specifically in the intestine. Therefore, we settled on the functional analysis of bacterial contigs of the shotgun library.

The *H. medicinalis* metagenome contains genes encoding proteins involved in the biosynthesis and transport of water-soluble vitamins and coenzymes that are known to be deficient in blood. In particular, we identified genes involved in the biosynthesis of thiamin (vitamin B1), riboflavin (B2), pantothenate (B5), pyridoxine (B6), biotin (B7), folate (B9) and cyanocobalamin (B12) (**Supplementary Figs. 18-24**). This finding is in good agreement with previous studies[26] describing the medicinal leech metagenome and confirms the suggestion that the digestive tract microbiota contributes to the biosynthesis of vitamins that the medicinal leech is incapable of obtaining in sufficient amounts from its blood meal or synthesizing by itself.

Blood digestion in blood-feeding animals is associated with haem and excess iron toxicity. Therefore, the leech digestive tract microbiota should as well be adapted to these toxicities. As expected, we identified a variety of microbial genes involved in iron acquisition and haem utilization. The identified genes encode ABC transporters for divalent iron, magnesium, and zinc cations, trivalent iron cations, and molecules that are capable to form stable iron complexes such as siderophores, haem or vitamin B12 (**Supplementary Fig.25**).

Unexpectedly, in the metagenomes of leeches collected from natural sources, we identified not only a variety of genes encoding bacterial multidrug efflux pumps but also clusters of genes that provide specific antibiotic resistance to tetracyclines, macrolides, fluoroquinolones, and beta lactams. We also found genes conferring resistance to cationic antimicrobial peptides (CAMPs). These genes encode bacterial two-component systems, such as BasS–R, ParS-R, PhoP-Q, protease ZapA and form the arnBCADTEF operon (**Supplementary Fig.26**). The CAMPs appear to be produced by the leeches, and

their interactions with bacterial two-component systems may play an important role in natural selection on the microbiome and in the regulation of the microbial community in the digestive system.

Discussion

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[30–34] is of great importance for public health care, veterinary medicine, and microbiology.

On the other hand, during the course of evolution, haematophagous organisms have developed strategies to counteract host defence mechanisms, such as blood coagulation, inflammation, immune response, and vasoconstriction at the site of injury. Although many haematophagous species are evolutionarily distant, the saliva components that mediate a successful blood meal are similar and have the same targets in their hosts. Evidently, this similarity results from convergence; however, each species possesses unique components.

Blood-sucking leeches are the only haematophagous organisms that are of use in medicine. Leeches of the genus *Hirudo* are most commonly used for medical purposes and include several closely related species, which are known to undergo cross-species hybridization. The genome sequencing of the medicinal leech fills a gap in understanding the genetic basis for the adaptation of the leech to blood feeding. Previously, the transcriptome of the leech central nervous system was annotated[22], the expression of innate immunity components was examined[35], and transcriptional analysis of the salivary glands of jawless and jawed leeches was performed[36–39].

Notably, the salivary cells of jawed leeches, as opposed to those of jawless leeches, do not form a discrete gland. Their cell bodies are interspersed between muscle fibres, and each salivary cell extends a single duct from the cell body to the jaw[7,40]. Due to these anatomical features, the real gene expression profile of salivary cells may not be accurately reflected by transcriptomes obtained using RNA extracts from rostral segments of the leech with subsequent cDNA sequencing by a relatively low-throughput EST approach[37,38,41]. Therefore, we used laser microdissection to obtain the salivary cell transcriptomes of three closely related leech species, which allowed us to analyse the differential gene expression in salivary cells with respect to that in adjacent muscle tissues.

To increase the accuracy of proteomic analysis, we improved the procedure for collecting SCS to minimize contamination. To cover the broadest repertoire of salivary molecules, we applied different mass spectrometry techniques to the samples obtained by various preparation methods.

Integration of the transcriptomic and proteomic analyses demonstrates that SCS of three leech species have a similar composition. SCS-specific proteins come from several main groups: enzymes, protease inhibitors, proteins involved in adhesion, and auxiliary proteins, which have closely related functions, such as hemostatic, antimicrobial, and digestive. 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. 27-47.**

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[42]. These enzymes are often found in snake venoms while the transcripts are observed in sialotranscriptomes of various hematophagous species[43–45]. In haemostasis, secreted proteases of the M12 family can participate in the inhibition of platelet adhesion[46,47] 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[48]. 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*)[49,50].

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[51]. 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[52]. Notably, leech salivary gland secretions contain carboxypeptidase inhibitors, which presumably prevent the untimely digestion of the blood meal by other types of peptidases[7,53].

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[54]. Interestingly, peroxynitrite may suppress haemostatic function by the nitration of key procoagulants[54,55], 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[56]. 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[57].

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[58,59]. In haematophagous animals, this enzyme can maintain optimal conditions for the digestion of a blood meal[60,61]. 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)[62]. 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[63]. In addition, the low-molecular-weight heparin produced by cleavage suppresses and inhibits blood coagulation[64].

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[65]. Apyrases are anticoagulants because they remove ADP, an important inducer of platelet aggregation at sites of tissue

injury[66].

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[67]. ADA is also found in the salivary gland secretion of the vampire snail *C. reticulata*, which belongs to *Spiralia,* as well as leeches[49]. ADA is thought to play an important role in the removal of adenosine because of its involvement in pain perception processes[68] .

Proteinase inhibitors

Antistasins: We identified sequences corresponding to proteinase inhibitor I15 (leech antistasin). 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[7]. Decorsin, an antistasin from *Macrobdella decora,* was demonstrated to inhibit platelet aggregation[69], and gigastasin from the giant Amazon leech (*Hementaria ghilianii*) was recently reported to potently inhibit complement C1[70]. Antistasin from *Hementeria officinalis* is the closest homologue of the sequences identified in our study.

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). They are commonly found in the venoms of snakes and other reptiles, and most of them are toxins[71,72]. 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[73] and the possible platelet aggregation inhibitor HPI[74], as well as to the snake toxins triflin (*Protobothrops flavoviridis*) and natrin-1 (*Naja atra*)[75,76].

Eglin-like: Eglins are small cysteine-free proteins that belong to the I13 family of serine proteinase inhibitors[77]. 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[7]. Of note, sequences identified in the present study have low homology to the classical eglin from leech.

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)[78] and are often found in the sialotranscriptomes of various ticks[79]. 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[80]. The nematode *Nippostrongylus brasiliensis* utilizes cystatins to evade the host immune system[81].

PAN domain: This domain is present in numerous proteins, including the blood proteins plasminogen and coagulation factor XI[82]. 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[83]. 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[84]. This protein exhibits affinity for collagens I, III, and IV and thereby inhibits collagen-mediated platelet adhesion[84].

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[85]. It can act as a plasmin inhibitor, thereby inhibiting fibrinolysis, but in some cases, it inhibits coagulation by inactivating thrombin and kallikrein[86]. 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[87]. In invertebrates, ficolins are involved in the recognition of bacterial cell wall components[88]. 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[89]. In reptile venoms, ficolin-like proteins, ryncolin[90] (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[91]. The DS domain plays an important role in the binding of various ligand molecules, including phospholipids and carbohydrates[92]. Due to these features, DS-containing proteins are actively involved in cell adhesion, migration, and the proliferation and activation of signalling cascades[93]. 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[92] 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[94]. 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[95]. 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[96]. They are present in scavenger receptors (mannose, fucose, collagen receptors), N-acetylgalactosaminyltransferases, haemolytic toxins (CEL-III from *Cucumaria echinata*) and apoptosis-inducing cytotoxins[96,97]. Previously, similar sequences were identified in leech transcriptomes; however, the authors assumed that this molecule has a mitochondrial localization[41]. 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[98]. As many known R-type lectins are involved in adhesion and trigger haemolysis[96], 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[99]. One protein identified in the leech proteome is a secreted protein that consists of four copies of the *vWFA* domain. 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)[49].

Accessory proteins and proteins with unknown functions

vWFD domain: We identified sequences that encode parts of a larger protein, which is the product of a single gene (g10240.t1) and a large, complex protein that comprises several domains, such as the von

Willebrand factor type D domain, C8 domain, and trypsin inhibitor-like Cys-rich domain. The combination of all such domains within a single protein features spondins and mucins[100]. Most secreted mucins are major components of mucus and mucosa and are found in salivary gland secretions, where they perform diverse functions[101].

Calreticulin: This protein family has various functions and is involved in diverse processes. The intracellular functions include chaperon activity, and these proteins play an important role in embryogenesis[102,103]. A number of studies have shown that calreticulin plays an active role in haemostasis and interacts with numerous coagulation and endothelial factors[104]. *In vivo* experiments revealed that these proteins exhibit antithrombotic activity[105]. Due to this property, calreticulin has been patented as an anticoagulant. This protein family is often found in omics data for haematophagous species[106,107]. They may play a modulating role in host haemostasis through binding calcium ions, which act as coagulation cofactors. Leech calreticulin was demonstrated to interact with the complement component C1q and plays an important role in restoring normal synaptic connections in the central nervous system of the leech[108].

Calmodulin: Calmodulin is a conserved multifunctional calcium binding protein containing the EF-hand domain. This protein was shown to activate intracellular eNOS in endothelial cells and to regulate collagen-dependent platelet aggregation[109]. Calmodulin, like calreticulin, is presumed to participate in the removal of free calcium ions.

Thioredoxin-like domain containing protein: Thioredoxin transfers electrons, acts as a chaperone, and, when coupled with glutathione, is a strong antioxidant involved in reactive oxygen species scavenging[110]. This protein has been found to play an important role in the immune response that emerged from peptide antigen–class I MHC interactions[111] and to participate in platelet activation and regulation[112]. Human proteins homologous to thioredoxin (for example, disulfide-isomerase A6) play a role in platelet aggregation induced by agonists, such as convulxin, collagen, and thrombin[113]. Some human extracellular disulfide isomerases are known to mediate platelet aggregation and thrombus formation (ERp57)[114]. They are secreted by platelets on the surface of the membrane and interact with calnexin and calreticulin.

In addition to the listed proteins, we identified globin, EGF-like protein, arginine ADP-ribosyltransferase, aspartyl aminopeptidase, and aldehyde dehydrogenase. The role of these molecules in haemostasis is currently unclear. In both proteomic and transcriptomic data, we identified cytoskeletal proteins, transcription factors, heat shock proteins, and ribosomal proteins.

Many of the identified components may play a role (known or putative) in anticoagulation and interactions with the host organism. Therefore, we propose a tentative scheme to explain the interactions between salivary proteins and components of hemostasis (**Fig. 6**). For a long time, thrombin and factor Xa have been thought to be the main targets of components of leech SCSs. However, similar to those of other haematophagous species, the SCSs of medicinal leeches were found to contain a complex mixture of antihaemostatic molecules aimed at inhibiting secondary and primary haemostasis. We suppose that the action of the SCS components during different stages of haemostasis and not merely on the key coagulation factors (thrombin and factor Xa) provides synergistic effects.

During the feeding process, leeches attach themselves to their host over a long period of time and optimally remain unnoticed by the host and its immune system. Blood feeding by leeches may last for an hour, which is sufficient for the appearance of the host immune response. Therefore, in addition to inhibiting blood coagulation, the SCS should suppress inflammation and the host immune response. SCS components are thought not only to facilitate the feeding process but also to inhibit blood coagulation factors and immune response proteins to prevent coagulation and quick proteolysis of the blood after ingestion[7].

In the SCSs of the medicinal leeches, we identified proteins homologous to others found in a variety of

species, including vertebrates. 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, to avoid interactions with components of the host immune system, appear to affect the kinetics of biochemical reactions via their homology, thereby providing a synergistic effect of the SCS.

Unexpectedly, we found that genes encoding well-known components of the SCS, such as hirudin, bdellins, eglins, saratin, and destabilase, were not differentially expressed in the salivary cells. Apparently, the proteins encoded by these genes perform some functions, in addition to blood feeding.

In the proteome of the SCS, 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[115], the mode of the secretion process in jawed leeches remains unknown. The presence of proteins with cytoplasmic or membrane localization appears to be associated with apocrine secretion in the production of leech saliva. Apart from the identified proteins, we detected previously unknown salivary proteins, the roles and importance of which in the leech have yet to be established. Therefore, functional analysis of the identified proteins demonstrated that leeches utilize ancient, highly conserved molecules as universal affectors of the host haemostasis and immunity.

The metagenomics of haematophagous species has attracted considerable attention to gain insight into the role of the microbiome in such species, which have a rather unusual way of feeding, to determine whether there are symbionts with unique properties, and to identify strategies used by pathogens for the colonization of the digestive tracts of their vectors/hosts. The microbiota of leeches has been extensively studied by microscopy, culturing, 16S rRNA gene sequencing, and imaging using FISH[26,116–119]. In contrast to previous studies, we combined laser microdissection, sequencing, and bioinformatics to elucidate the composition of bacterial communities and performed functional characterization in morphologically and functionally different compartments of the leech digestive tract.

We found that despite the different geographic origin of the leeches, their microbiomes have a very similar bacterial composition. The crop and intestine of each leech species have a similar bacterial composition that was in good agreement with the findings of previous studies at high taxonomic ranks[26,119]. Moreover, the results of 16S rRNA gene sequencing correlate with the binning of metagenomic contigs in the *H. medicinalis* shotgun library. The correlation between the results obtained by independent methods supports the significance of our metagenomic data and demonstrates that the microbiome of the entire leech comprises mainly the bacterial community of the crop, the largest compartment of the digestive tract.

The microbiomes of leeches contain bacterial taxa characteristic of these species. We identified several bacterial genera (*Mucinivorans*, *Aeromonas, Niabella)* that are closely associated with the leech organism and are cultivable, some of which have sequenced genomes[24,25,120]. We also found bacterial taxa that have not been described previously in the microbiomes of leeches but have been identified in other haematophagous species, *e.g*., those of the genus *Burkholderia*, which are predominant in the crop.

In contrast to previous studies, in which only the bacterial composition was explored, we performed metagenomic analysis of the digestive tract compartments to evaluate the overall metabolic potential of the leech microbiota and determine the contributions of the individual microbiome members to the metabolic pathways involved in blood storage and digestion. These data may be used to compare not only blood-sucking leech metagenomes but also the metagenomes of haematophagous species from various taxa. Comparative analyses of the metagenomes of haematophagous species will provide insight into the role of the microbiota in the adaptation to blood feeding.

Finally, the detection of antibiotic resistance genes in the metagenome of the leech is attributable to the

use of antibiotics in the agriculture industry. Antibiotic-resistant bacteria can be acquired by the leech both directly during feeding on farm animals and from water environments (rivers and water reservoirs). Therefore, the principles of leech farming for hirudotherapy should be reviewed regarding the risk of antibiotic-resistant infections.

As described above, in this paper, we annotate the *de novo* assembled genome of *H. medicinalis*. This genome annotation offers a roadmap for future experimentation on the medicinal leech as a model organism and provides a database of sequences encoding the unique leech proteins for use in developing pharmacological compounds. Identifying previously unknown homologues of the genes encoding leech anticoagulants and other known blood feeding-related proteins in the annotated genome could provide 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. Differential expression analysis validated by proteomic data identified the genes uniquely expressed in the salivary cells many of them were described for the first time as genes encoding components of the SCS. Functional analysis of the identified genes enabled the reconstruction of pathways by which the SCS proteins might interact with the haemostatic components. Finally, the metagenome data analysis not only identified the taxonomic composition of the crop and intestine microbiota but also revealed its overall metabolic potential. The detection of antibiotic resistance genes in the metagenomes of the leeches collected from natural sources allows us to suggest the direct impact of environmental and anthropogenic factors (agricultural industry) on the leech microbiota, which is critical for leech farming and subsequent clinical use.

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[9] was used to create a single genome assembly. For eukaryotic contig scaffolding, we used Sspace software[13] 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[122] . 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[123]. The depth of read coverage for the individual libraries was calculated using BEDTools[124]. The taxonomic classification of the contigs was carried out by MEGAN6[10] 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[125] (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[126] v1.9 was applied to trim the adapter sequences used for cDNA synthesis. Prinseq lite[127] v.0.20.4 was used to trim reads according to their quality and length. For *de novo* transcriptome assembly, we used Trinity software[18] (version r20131110) with the default parameters. Blast2GO[19] 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[10] 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[123], Htseq[128] and edgeR[129] 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[130]. 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[128] 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.

Illumina 16S rRNA sequencing for metagenomic analysis. We extracted the total DNA from tissue fragments isolated by laser microdissection using a QIAamp DNA Micro Kit (Qiagen, Germany) according to the 'Isolation of Genomic DNA from Laser-Microdissected Tissues' protocol with slight modifications (for details, see **Supplementary Methods**). The 16S rRNA libraries were constructed by the amplification of V3 and V4 hypervariable regions by consecutive application of the Nextera XT Index Kit (Illumina) with a dual indexing strategy according to the '16S metagenomic sequencing library preparation' protocol (Illumina, San Diego, CA, USA). The 16S rRNA libraries were sequenced on the MiSeq sequencing platform (Illumina) using a 2 × 250 Cycle V2 Kit according to the standard Illumina sequencing protocols. For 16S rRNA analysis, we used the greengenes database (version gg_13_5) and the MEGAN6 software[10].

References.

1. Sket, B. & Trontelj, P. Global diversity of leeches (Hirudinea) in freshwater. *Hydrobiologia* **595,** 129–137 (2008).

2. Wagenaar, D. A. A classic model animal in the 21st century: recent lessons from the leech nervous system. *J. Exp. Biol.* **218,** 3353–9 (2015).

3. Rados, C. Beyond bloodletting: FDA gives leeches a medical makeover. *FDA Consum.* **38,** 9

4. Hyson, J. M. Leech therapy: a history. *J. Hist. Dent.* **53,** 25–7 (2005).

5. Houschyar, K. S. *et al.* Medical leech therapy in plastic reconstructive surgery. *Wiener Medizinische Wochenschrift* **165,** 419–425 (2015).

6. Basanova, A. V, Baskova, I. P. & Zavalova, L. L. Vascular-platelet and plasma hemostasis regulators from bloodsucking animals. *Biochemistry. (Mosc).* **67,** 143–50 (2002).

7. Hildebrandt, J. & Lemke, S. Small bite, large impact – saliva and salivary molecules in the medicinal leech, Hirudo medicinalis. *Naturwissenschaften* **165,** 995–1008 (2011).

8. Rigbi, M. *et al.* The saliva of the medicinal leech Hirudo medicinalis--I. Biochemical characterization of the high molecular weight fraction. *Comp. Biochem. Physiol. B.* **87,** 567–73 (1987).

9. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–77 (2012).

10. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–86 (2007).

11. Sedlar, K., Kupkova, K. & Provaznik, I. Bioinformatics strategies for taxonomy independent binning and visualization of sequences in shotgun metagenomics. *Comput. Struct. Biotechnol. J.* **15,** 48–55 (2017).

12. Nikitina, A. *et al.* Draft mitochondrial genomes of Hirudo medicinalis and Hirudo verbana (Annelida, Hirudinea). *Mitochondrial DNA Part B Resour.* **1,** 254–256 (2016).

13. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27,** 578–579 (2011).

14. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6,** 11 (2015).

15. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493,** 526–531 (2012).

16. Albertin, C. B. *et al.* The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature* **524,** 220–224 (2015).

17. Emms, D. M. & Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **16,** 157 (2015).

18. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–52 (2011).

19. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676 (2005).

20. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44,** D286–D293 (2016).

21. Anders, S. *et al.* Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8,** 1765–1786 (2013).

22. Hibsh, D., Schori, H., Efroni, S. & Shefi, O. De novo transcriptome assembly databases for the central nervous system of the medicinal leech. *Sci. Data* **2,** 150015 (2015).

23. Baskova, I. P., Yudina, T. G., Zavalova, L. L. & Dudkina, A. S. Protein-lipid particles of medicinal leech salivary gland secretion; their size and morphology. *Biochemistry. (Mosc).* **75,** 585–9 (2010).

24. Glaeser, S. P., Galatis, H., Martin, K. & Kämpfer, P. Niabella hirudinis and Niabella drilacis sp. nov., isolated from the medicinal leech Hirudo verbana. *Int. J. Syst. Evol. Microbiol.* **63,** 3487–3493 (2013).

25. Nelson, M. C., Bomar, L., Maltz, M. & Graf, J. Mucinivorans hirudinis gen. Nov., sp. nov., an anaerobic, mucin-degrading bacterium isolated from the digestive tract of the medicinal leech Hirudo verbena. *Int. J. Syst. Evol. Microbiol.* **65,** 990–995 (2015).

26. Maltz, M. A. *et al.* Metagenomic analysis of the medicinal leech gut microbiota. *Front. Microbiol.* **5,** 151 (2014).

27. Worthen, P. L., Gode, C. J. & Graf, J. Culture-independent characterization of the digestive-tract microbiota of the medicinal leech reveals a tripartite symbiosis. *Appl. Environ. Microbiol.* **72,** 4775–81 (2006).

28. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

29. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* **428,** 726–731 (2016).

30. Chen, X.-G. *et al.* Genome sequence of the Asian Tiger mosquito, Aedes albopictus , reveals insights into its biology, genetics, and evolution. *Proc. Natl. Acad. Sci.* **112,** E5907–E5915 (2015).

31. Mesquita, R. D. *et al.* Genome of Rhodnius prolixus , an insect vector of Chagas disease, reveals unique adaptations to hematophagy and parasite infection. *Proc. Natl. Acad. Sci.* **112,** 14936–14941 (2015).

32. Zepeda Mendoza, M. L. *et al.* Hologenomic adaptations underlying the evolution of sanguivory in the common vampire bat. *Nat. Ecol. Evol.* **2,** (2018).

33. Gulia-nuss, M. *et al.* Genomic insights into the Ixodes scapularis tick vector of Lyme disease.*Nat. Commun.* **7,** 10507 (2016).

34. Rosenfeld, J. A. *et al.* Genome assembly and geospatial phylogenomics of the bed bug Cimex lectularius. *Nat. Commun.* **7,** 10164 (2016).

35. Macagno, E. R. *et al.* Construction of a medicinal leech transcriptome database and its application to the identification of leech homologs of neural and innate immune genes. *BMC Genomics* **11,** 407 (2010).

36. Siddall, M. E., Brugler, M. R. & Kvist, S. Comparative Transcriptomic Analyses of Three Species of Placobdella (Rhynchobdellida: Glossiphoniidae) Confirms a Single Origin of Blood Feeding in Leeches. *J. Parasitol.* **102,** 143–150 (2016).

37. Kvist, S., Min, G.-S. & Siddall, M. E. Diversity and selective pressures of anticoagulants in three medicinal leeches (Hirudinida: Hirudinidae, Macrobdellidae). *Ecol. Evol.* **3,** 918 (2013).

38. Kvist, S., Brugler, M. R., Goh, T. G., Giribet, G. & Siddall, M. E. Pyrosequencing the salivary transcriptome of Haemadipsa interrupta (Annelida: Clitellata: Haemadipsidae): anticoagulant diversity and insight into the evolution of anticoagulation capabilities in leeches. *Invertebr. Biol.* **133,** 74–98 (2014).

39. Amorim, A. M. X. P. *et al.* Transcripts involved in hemostasis: Exploring salivary complexes from Haementeria vizottoi leeches through transcriptomics, phylogenetic studies and structural features. *Toxicon* **106,** 20–29 (2015).

40. 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* **8,** e73809 (2013).

41. Min, G.-S., Sarkar, I. N. & Siddall, M. E. Salivary Transcriptome of the North American Medicinal Leech, Macrobdella decora. *J. Parasitol.* **96,** 1211–1221 (2010).

42. Seals, D. F. & Courtneidge, S. A. The ADAMs family of metalloproteases: multidomain proteins with multiple functions. *Genes Dev.* **17,** 7–30 (2003).

43. Kini, R. & Koh, C. Metalloproteases Affecting Blood Coagulation, Fibrinolysis and Platelet Aggregation from Snake Venoms: Definition and Nomenclature of Interaction Sites. *Toxins (Basel).* **8,** 284 (2016).

44. Chagas, A. C. *et al.* A deep insight into the sialotranscriptome of the mosquito, Psorophora albipes. *BMC Genomics* **14,** 875 (2013).

45. Ribeiro, J. M. C., Slovák, M. & Francischetti, I. M. B. An insight into the sialome of Hyalomma excavatum. *Ticks Tick. Borne. Dis.* **8,** 201–207 (2017).

46. Kini, R. M. Anticoagulant proteins from snake venoms: structure, function and mechanism. *Biochem. J.* **397,** 377–87 (2006).

47. Brahma, R. K., McCleary, R. J. R., Kini, R. M. & Doley, R. Venom gland transcriptomics for identifying, cataloging, and characterizing venom proteins in snakes. *Toxicon* **93,** 1–10 (2015).

48. Bland, N. D., Pinney, J. W., Thomas, J. E., Turner, A. J. & Isaac, R. E. Bioinformatic analysis of the neprilysin (M13) family of peptidases reveals complex evolutionary and functional relationships. *BMC Evol. Biol.* **8,** 16 (2008).

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

50. Rivière, G. *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.* **382,** 565–73 (2004).

51. Gingras, R. *et al.* 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.* **274,** 11742–50 (1999).

52. Sojka, D. *et al.* New insights into the machinery of blood digestion by ticks. *Trends Parasitol.* **29,** 276–85 (2013).

53. Reverter, D. *et al.* A carboxypeptidase inhibitor from the medical leech Hirudo medicinalis. Isolation, sequence analysis, cDNA cloning, recombinant expression, and characterization. *J. Biol. Chem.* **273,** 32927–33 (1998).

54. Nielsen, V. G., Crow, J. P., Mogal, A., Zhou, F. & Parks, D. A. Peroxynitrite decreases hemostasis in human plasma in vitro. *Anesth. Analg.* **99,** 21–6 (2004).

55. Kurahashi, T. & Fujii, J. Roles of Antioxidative Enzymes in Wound Healing. *J. Dev. Biol. 2015, Vol. 3, Pages 57-70* **3,** 57–70 (2015).

56. Crispell, G., Budachetri, K. & Karim, S. Rickettsia parkeri colonization in Amblyomma maculatum: The role of superoxide dismutases. *Parasites and Vectors* **9,** 1–12 (2016).

57. Graça-Souza, A. V *et al.* Adaptations against heme toxicity in blood-feeding arthropods. *Insect Biochem. Mol. Biol.* **36,** 322–35 (2006).

58. Tripp, B. C., Smith, K. & Ferry, J. G. Carbonic anhydrase: new insights for an ancient enzyme. *J. Biol. Chem.* **276,** 48615–8 (2001).

59. Carter, M. J. & Parsons, D. S. The isoenzymes of carbonic anhydrase: tissue, subcellular distribution and functional significance, with particular reference to the intestinal tract. *J. Physiol.* **215,** 71–94 (1971).

60. Santos, V. C., Nunes, C. A., Pereira, M. H. & Gontijo, N. F. Mechanisms of pH control in the midgut of Lutzomyia longipalpis: roles for ingested molecules and hormones. *J. Exp. Biol.* **214,** 1411–8 (2011).

61. Linser, P. J., Smith, K. E., Seron, T. J. & Neira Oviedo, M. Carbonic anhydrases and anion transport in mosquito midgut pH regulation. *J. Exp. Biol.* **212,** 1662–71 (2009).

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

63. Kemparaju, K. & Girish, K. S. Snake venom hyaluronidase: a therapeutic target. *Cell Biochem. Funct.* **24,** 7–12 (2006).

64. Shriver, Z. *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.* **97,** 10365–70 (2000).

65. Rigbi, M., Orevi, M. & Eldor, A. Platelet aggregation and coagulation inhibitors in leech saliva and their roles in leech therapy. *Semin. Thromb. Hemost.* **22,** 273–8 (1996).

66. Yip, J., Shen, Y., Berndt, M. C. & Andrews, R. K. Primary platelet adhesion receptors. *IUBMB Life* **57,** 103–108 (2005).

67. Ribeiro, J. M., Charlab, R. & Valenzuela, J. G. The salivary adenosine deaminase activity of the mosquitoes Culex quinquefasciatus and Aedes aegypti. *J. Exp. Biol.* **204,** 2001–10 (2001).

68. Burnstock, G. & Wood, J. N. Purinergic receptors: Their role in nociception and primary afferent neurotransmission. *Curr. Opin. Neurobiol.* **6,** 526–532 (1996).

69. Seymour, J. L., Henzel, W. J., Nevins, B., Stults, J. T. & Lazarus, R. A. Decorsin. A potent glycoprotein IIb-IIIa antagonist and platelet aggregation inhibitor from the leech Macrobdella decora. *J. Biol. Chem.* **265,** 10143–7 (1990).

70. Pang, S. S. *et al.* The Structural Basis for Complement Inhibition by Gigastasin, a Protease Inhibitor from the Giant Amazon Leech. *J. Immunol.* **199,** 3883–3891 (2017).

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

72. Koludarov, I. *et al.* Fossilized venom: the unusually conserved venom profiles of Heloderma species (beaded lizards and gila monsters). *Toxins (Basel).* **6,** 3582–95 (2014).

73. Chhabra, S. *et al.* Kv1.3 channel-blocking immunomodulatory peptides from parasitic worms: implications for autoimmune diseases. *FASEB J.* **28,** 3952–64 (2014).

74. Ma, D., Francischetti, I. M. B., Ribeiro, J. M. C. & Andersen, J. F. The structure of hookworm platelet inhibitor (HPI), a CAP superfamily member from Ancylostoma caninum. *Acta Crystallogr. Sect. F, Struct. Biol. Commun.* **71,** 643–9 (2015).

75. 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.* **350,** 735–43 (2005).

76. Wang, Y.-L. *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.* **285,** 37872–83 (2010).

77. Rawlings, N. D., Tolle, D. P. & Barrett, A. J. Evolutionary families of peptidase inhibitors. *Biochem. J.* **378,** 705–16 (2004).

78. Grzonka, Z. *et al.* Structural studies of cysteine proteases and their inhibitors. *Acta Biochim. Pol.* **48,** 1–20 (2001).

79. Wang, Y. *et al.* Functional characterization of a cystatin from the tick Rhipicephalus haemaphysaloides. *Parasit. Vectors* **8,** 140 (2015).

80. 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.* **7,** 216 (2017).

81. Dainichi, T. *et al.* Nippocystatin, a cysteine protease inhibitor from Nippostrongylus brasiliensis, inhibits antigen processing and modulates antigen-specific immune response. *Infect. Immun.* **69,** 7380–6 (2001).

82. Ho, D. H., Badellino, K., Baglia, F. A. & Walsh, P. N. A binding site for heparin in the apple 3 domain of factor XI. *J. Biol. Chem.* **273,** 16382–90 (1998).

83. Brown, P. J., Gill, A. C., Nugent, P. G., McVey, J. H. & Tomley, F. M. 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.* **497,** 31–38 (2001).

84. Huizinga, E. G. *et al.* The structure of leech anti-platelet protein, an inhibitor of haemostasis. *Acta Crystallogr. D. Biol. Crystallogr.* **57,** 1071–8 (2001).

85. Rehman, A. A., Ahsan, H. & Khan, F. H. α-2-Macroglobulin: a physiological guardian. *J. Cell. Physiol.* **228,** 1665–75 (2013).

86. 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.* **101,** 183–91 (2001).

87. Fujita, T. Evolution of the lectin–complement pathway and its role in innate immunity. *Nat. Rev. Immunol.* **2,** 346–353 (2002).

88. Angthong, P., Roytrakul, S., Jarayabhand, P. & Jiravanichpaisal, P. Characterization and function of a tachylectin 5-like immune molecule in Penaeus monodon. *Dev. Comp. Immunol.* **76,** 120–131 (2017).

89. Kairies, N. *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. U. S. A.* **98,** 13519–24 (2001).

90. OmPraba, G. *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.* **9,** 1882–1893 (2010).

91. Kane, W. H. & Davie, E. W. Blood coagulation factors V and VIII: structural and functional similarities and their relationship to hemorrhagic and thrombotic disorders. *Blood* **71,** 539–55 (1988).

92. Foster, P. A., Fulcher, C. A., Houghten, R. A. & Zimmerman, T. S. Synthetic factor VIII peptides with amino acid sequences contained within the C2 domain of factor VIII inhibit factor VIII binding to phosphatidylserine. *Blood* **75,** 1999–2004 (1990).

93. Leitinger, B. Discoidin Domain Receptor Functions in Physiological and Pathological Conditions. in *International review of cell and molecular biology* **310,** 39–87 (2014).

94. Jeon, H. & Blacklow, S. C. STRUCTURE AND PHYSIOLOGIC FUNCTION OF THE LOW-DENSITY LIPOPROTEIN RECEPTOR. *Annu. Rev. Biochem.* **74,** 535–562 (2005).

95. Zipser, B., Bradford, J. J. & Hollingsworth, R. I. Cholesterol and its derivatives, are the principal steroids isolated from the leech species Hirudo medicinalis. *Comp. Biochem. Physiol. C. Pharmacol. Toxicol. Endocrinol.* **120,** 269–82 (1998).

96. Gupta, R. K. & Gupta, G. S. R-Type Lectin Families. in *Animal Lectins: Form, Function and Clinical Applications* 313–330 (Springer Vienna, 2012). doi:10.1007/978-3-7091-1065-2_14

97. Uchida, T. *et al.* Crystal Structure of the Hemolytic Lectin CEL-III Isolated from the Marine Invertebrate Cucumaria echinata. *J. Biol. Chem.* **279,** 37133–37141 (2004).

98. Suzuki, R. *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.* **65,** 49–57 (2009).

99. Perkins, S. J. *et al.* 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.* **238,** 104–119 (1994).

100. Lang, T., Hansson, G. C. & Samuelsson, T. Gel-forming mucins appeared early in metazoan evolution. (2007).

101. Lang, T., Hansson, G. C. & Samuelsson, T. Gel-forming mucins appeared early in metazoan evolution. *Proc. Natl. Acad. Sci.* **104,** 16209–16214 (2007).

102. Park, B. J. *et al.* Calreticulin, a calcium-binding molecular chaperone, is required for stress response and fertility in Caenorhabditis elegans. *Mol. Biol. Cell* **12,** 2835–45 (2001).

103. Rauch, F., Prud'homme, J., Arabian, A., Dedhar, S. & St-Arnaud, R. Heart, Brain, and Body Wall Defects in Mice Lacking Calreticulin. *Exp. Cell Res.* **256,** 105–111 (2000).

104. Kuwabara, K. *et al.* Calreticulin, an antithrombotic agent which binds to vitamin K-dependent coagulation factors, stimulates endothelial nitric oxide production, and limits thrombosis in canine coronary arteries. *J. Biol. Chem.* **270,** 8179–87 (1995).

105. Dai, E. *et al.* Calreticulin, a potential vascular regulatory protein, reduces intimal hyperplasia after arterial injury. *Arterioscler. Thromb. Vasc. Biol.* **17,** 2359–68 (1997).

106. Kazimírová, M. & Štibrániová, I. Tick salivary compounds: their role in modulation of host defences and pathogen transmission. *Front. Cell. Infect. Microbiol.* **3,** 43 (2013).

107. Jaworski, D. C. *et al.* A secreted calreticulin protein in ixodid tick (Amblyomma americanum) saliva. *J. Insect Physiol.* **41,** 369–375 (1995).

108. Lefebvre, C. *et al.* Calreticulin contributes to C1q-dependent recruitment of microglia in the leech Hirudo medicinalis following a CNS injury. *Med. Sci. Monit.* **20,** 644–653 (2014).

109. Gardiner, E. E., Arthur, J. F., Berndt, M. C. & Andrews, R. K. Role of calmodulin in platelet receptor function. *Curr. Med. Chem. Cardiovasc. Hematol. Agents* **3,** 283–7 (2005).

110. Espinosa-Diez, C. *et al.* Antioxidant responses and cellular adjustments to oxidative stress. *Redox Biol.* **6,** 183–97 (2015).

111. Dick, T. B. Assembly of MHC class I peptide complexes from the perspective of disulfide bond formation. *Cell. Mol. Life Sci.* **61,** 547–556 (2004).

112. Mor-Cohen, R. Disulfide Bonds as Regulators of Integrin Function in Thrombosis and Hemostasis. *Antioxid. Redox Signal.* **24,** 16–31 (2016).

113. Jordan, P. A. *et al.* A role for the thiol isomerase protein ERP5 in platelet function. *Blood* **105,** 1500–1507 (2005).

114. Wu, Y. *et al.* The disulfide isomerase ERp57 mediates platelet aggregation, hemostasis, and thrombosis. *Blood* **119,** 1737–1746 (2012).

115. Walz, B., Schäffner, K.-H. & Sawyer, R. T. Ultrastructure of the anterior salivary gland cells of the giant leech, Haementeria ghilianii (Annelida, Hirudinea). *J. Morphol.* **196,** 321–332 (1988).

116. Perkins, S. L., Budinoff, R. B. & Siddall, M. E. New gammaproteobacteria associated with blood-feeding leeches and a broad phylogenetic analysis of leech endosymbionts. *Appl. Environ. Microbiol.* **71,** 5219–24 (2005).

117. Siddall, M. E., Perkins, S. L. & Desser, S. S. Leech mycetome endosymbionts are a new lineage of alphaproteobacteria related to the Rhizobiaceae. *Mol. Phylogenet. Evol.* **30,** 178–86 (2004).

118. Marden, J. N., McClure, E. A., Beka, L. & Graf, J. Host Matters: Medicinal Leech Digestive-Tract Symbionts and Their Pathogenic Potential. *Front. Microbiol.* **7,** 1569 (2016).

119. Ott, B. M., Rickards, A., Gehrke, L. & Rio, R. V. M. Characterization of shed medicinal leech mucus reveals a diverse microbiota. *Front. Microbiol.* **5,** 757 (2014).

120. Bomar, L. *et al.* Draft Genome Sequence of Aeromonas veronii Hm21, a Symbiotic Isolate from the Medicinal Leech Digestive Tract. *Genome Announc.* **1,** (2013).

121. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–7 (2000).

122. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–9 (2012).

123. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

124. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33,** W465-7 (2005).

125. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).

126. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).

127. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2015).

128. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).

129. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

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, transcriptome and 16S rRNA sequencing. V.V.B., A.S. Kasianov, and A.I.M. genome assembly and automated annotation. V.V.B., A.S. Kasianov, and D.V.V. transcriptome assembly and automated annotation. D.V.V., A.S.N. differential expression analysis. A.I.M, V.V.B. and M.T.V. binning and metagenome 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. D.D.K., N.F.P., P.A.B. contributed to sample preparation. 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., 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)

Figure legends

Fig. 1. The *H. medicinalis* genome. **(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)** Distribution of transposable elements. The inner chart represents the three main classes of TEs (LTRs, non-LTRs, and class II), and the outer one shows the distribution of TE superfamilies within each class. The charts are based on the total length of TE-related sequences. **(C)** Gene clustering using OrthoMCL software for Annelids (purple, *H. medicinalis*, *H. robusta*, *C. teleta*), Brichiopods (red, *L. anatine*), Mollusca (blue, *O. bimaculoides*, *A. californica*, *Crassostrea gigas*). Vertebrates (green, *D. rerio* and *H. sapiens*) serve as an outgroup.

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)**.

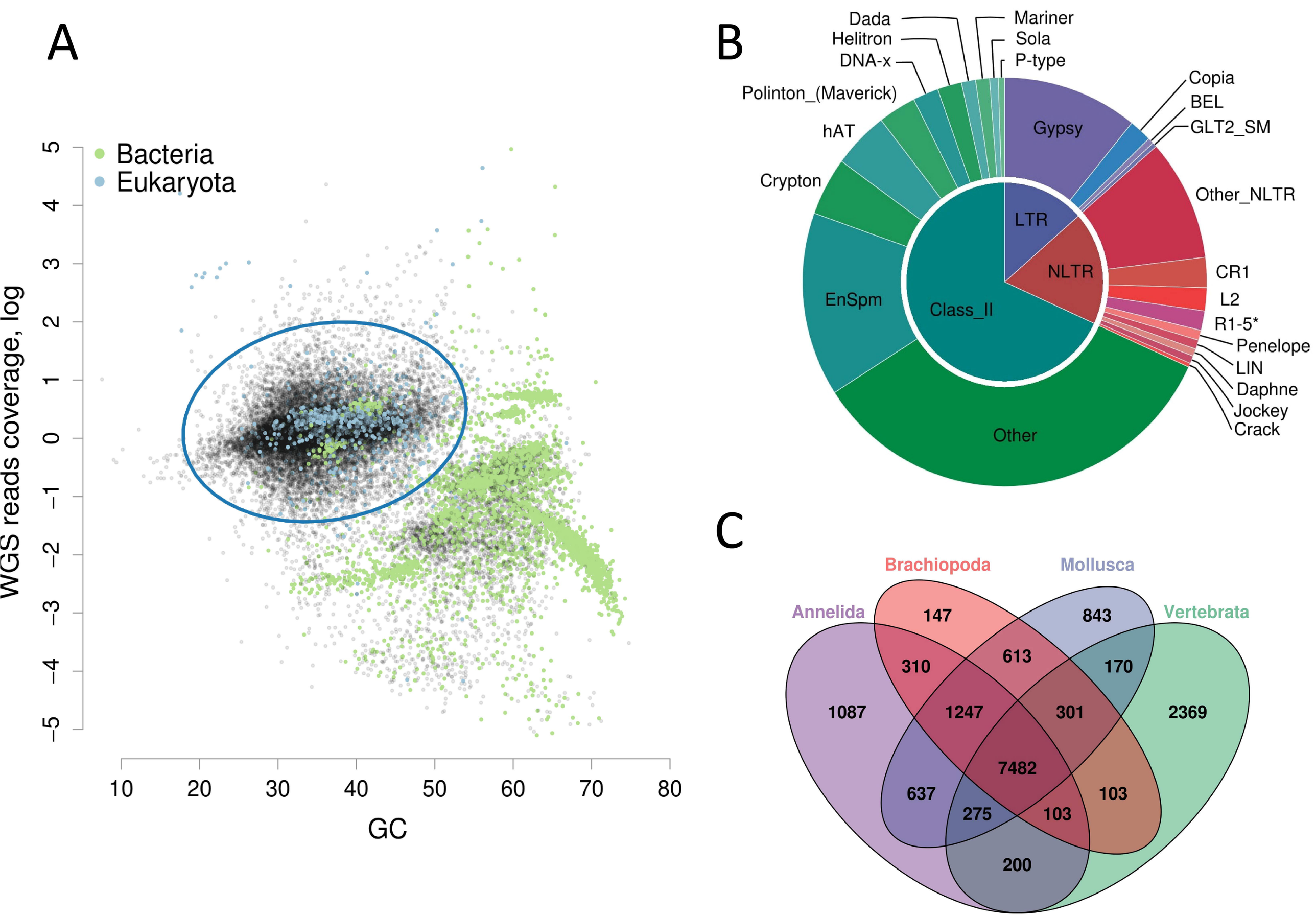
Fig. 3. Summary of the identified SCS components. The Venn diagrams in the upper panel show the numbers of ortholog clusters identified by differential expression (DE) and proteomic (Prot) analyses across three medicinal leech species. The histogram in the middle panel features the numbers of orthologous 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**.

Fig. 4. Metagenome of *H. medicinalis.* **(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 is encoded by colour (blue - *Burkholderiales,* soft orange - *Rhizobiales,* green - *Bacteroidales,* lime green - *Chitinophagales*, orange - *Sphingobacteriales,* red - *Myxococcales,* purple – *Gammaproteobacteria*, grey – other). Taxonomic affiliation was determined by direct Blastn search against the NCBI nt database **(B)** Beeswarm plots. Mean read coverage of contigs that belong to different bacterial taxa and *Eukaryota*.

Fig. 5. Principal component analysis of 16S rRNA sequencing data obtained for three medicinal leech species. The taxonomic composition in the intestine differs from that in the crop and coeca.

Fig. 6. Tentative scheme of interactions between the leech SCS proteins and components of host haemostasis. Movingitsjaws, leechbitesthrough the host body surface and destroys blood vessels that causes subendothelial collagen deposition into vessel lumen.Platelet adhesion is primarily mediated by direct interaction of collagen with glycoproteins (GP) (Ia-IIa, VI) or by binding platelet surface receptor (GP) Ib-IX-V complex to von Willebrand factor (vWF) in the subendothelial matrix. Adhesion of platelets causes their activation and releasing additional factors of homeostasis. Then conformational activation of the αIIbβ3 receptors occurs. Activated αIIbβ3 receptors bind to fibrinogen and vWF that leads to further aggregation of platelets. The leech SCS proteins (shown in red) interfere with the adhesion and activation of platelets. Proteases M12 participate in the cleavage of fibrinogen, vWF, and glycoproteins on the surface of platelets and the endothelium. Proteins containing the vWFA and F5/8 domains inhibit the interaction of the haemostatic factors with each other and collagen. Saratin, CRISP and PAN/apple inhibit the vWF interaction with collagen. Apyrases eliminate ADP, the main agonist of platelet

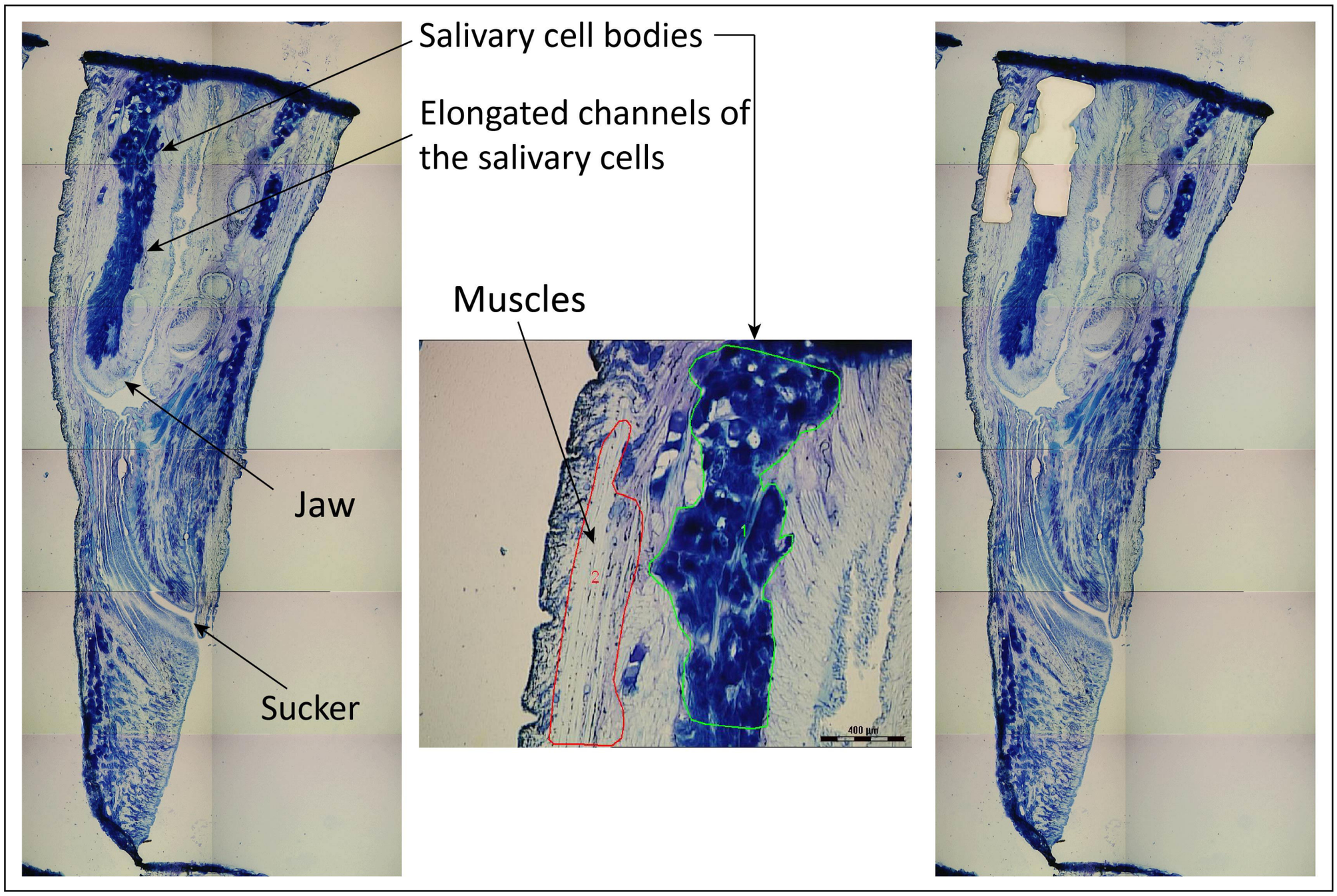
activation.The blood coagulation cascade is initiated at the injury site. The tissue factor (TF) binds to factor VII forming a protease complex that activates plasma factor X by its cleavage. Factor X binds to factor V to form an active protease complex that causes formation of thrombin from prothrombin. Thrombin is an additional inducer of platelets and activates them through the cleavage of the external domain of the protein-coupled protease-activated receptors (PAR). Thrombin is also able to cleave fibrinogen to insoluble fibrin. Activation of factor VII by collagen induces kallikrein and initiates the classical complement pathway. Antistasins are potential inhibitors of the majority of serine proteases involved in the cascade of the blood clotting. They mainly inhibit the formation of thrombin as well as the activation of complement pathway. Antistasins, a2M-like protein, and M13 protease (ACE) are probably involved in the regulation of blood pressure and prevent further inflammation. Along with destabilase, M12 proteases participate in the inhibition of thrombus formation. CRISP proteins, ADA, and SODC probably contribute analgesic effect of SCS.


A
Bacteria
Eukaryota
WGS reads coverage, log
GC
B
Dada
Helitron
DNA-x
Polinton_(Maverick)
hAT
Crypton
EnSpm
Mariner
Sola
P-type
Copia
BEL
GLT2_SM
Gypsy
Other_NLTR
LTR
NLTR
Class_II
CR1
L2
R1-5*
Penelope
LIN
Daphne
Jockey
Crack
Other
C
Brachiopoda
Mollusca
Annelida
Vertebrata
147
843
310
613
170
1087
1247
301
2369
7482
637
103
275
103
200

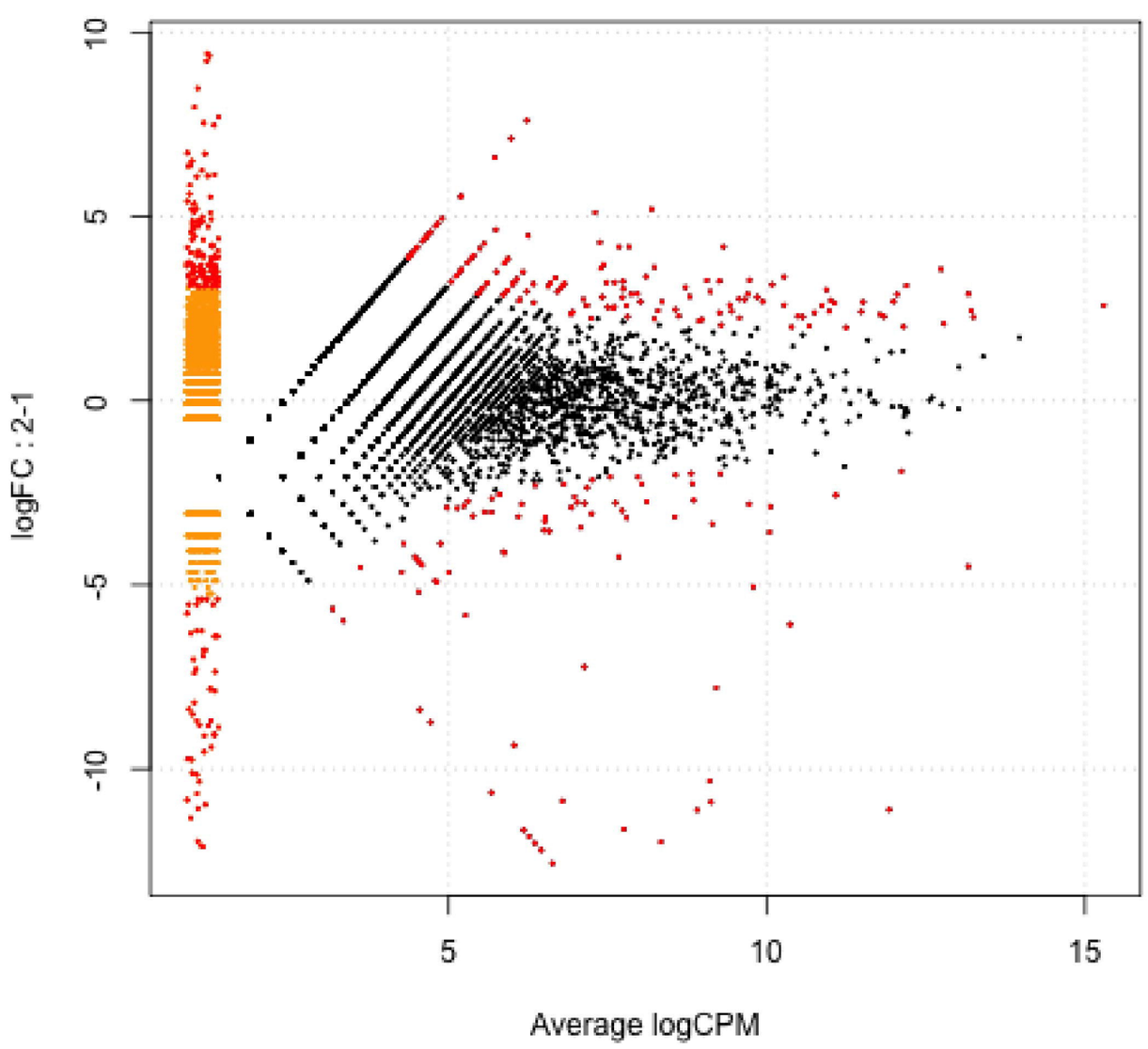
A

before LMD

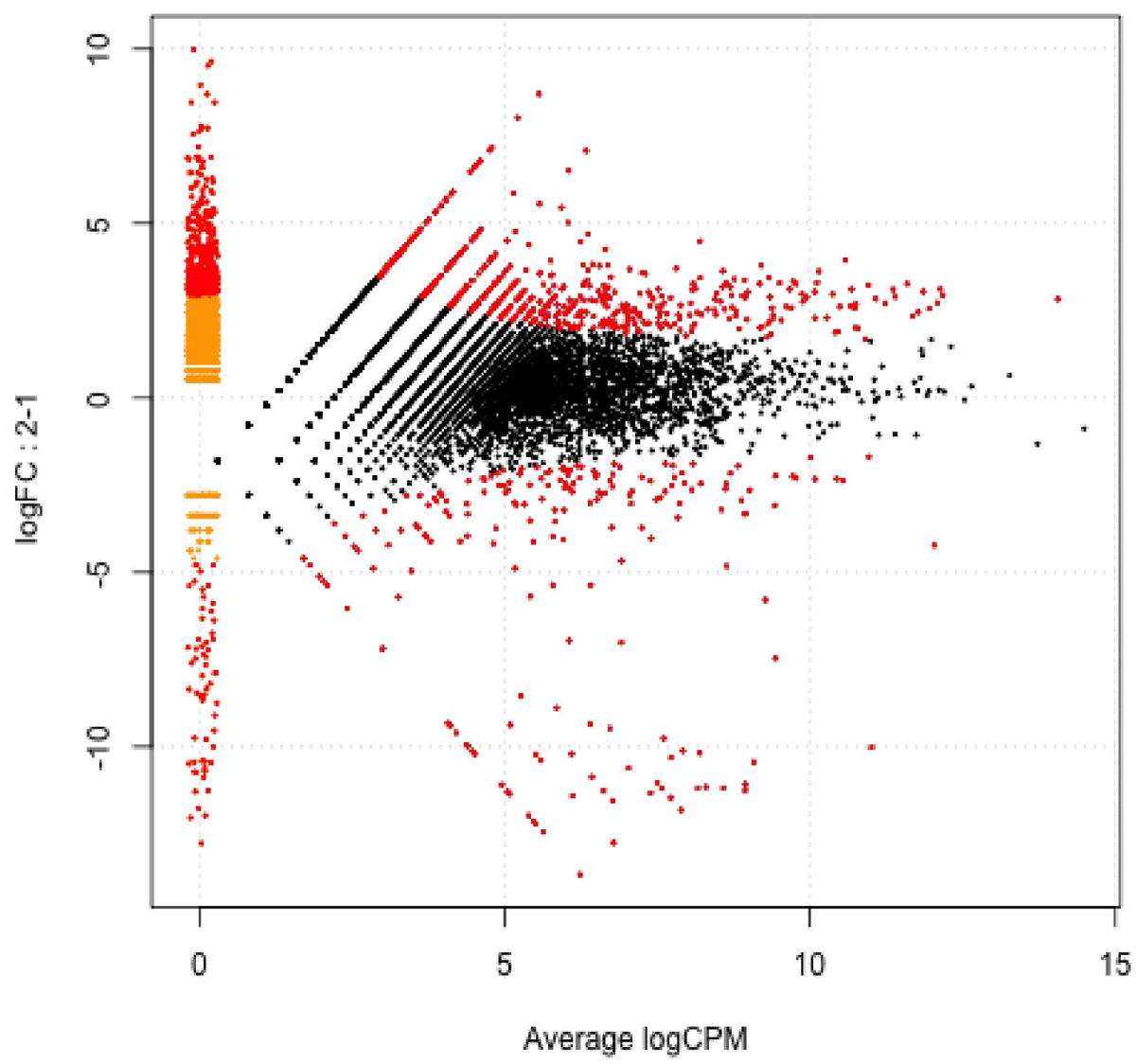
after LMD

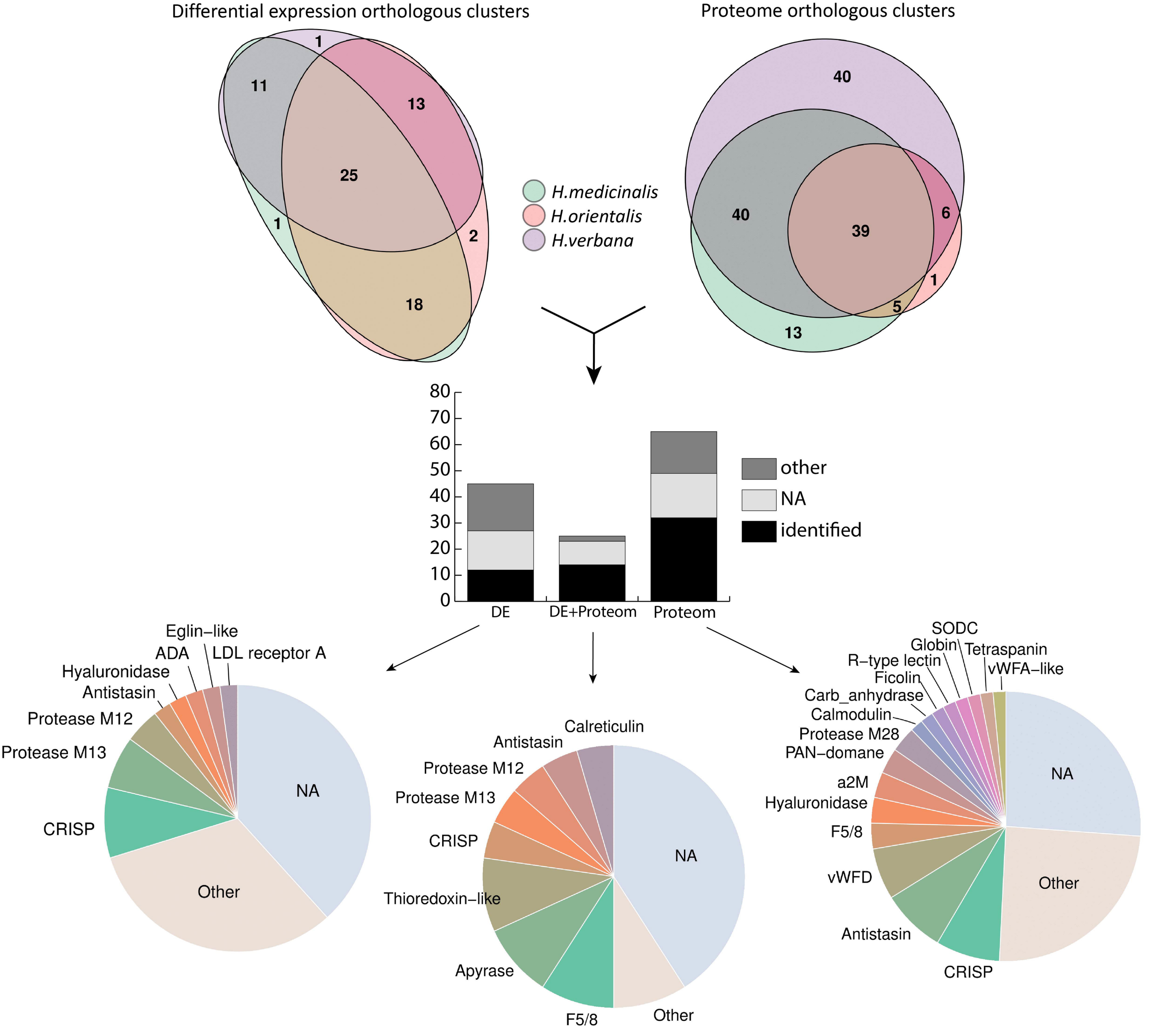


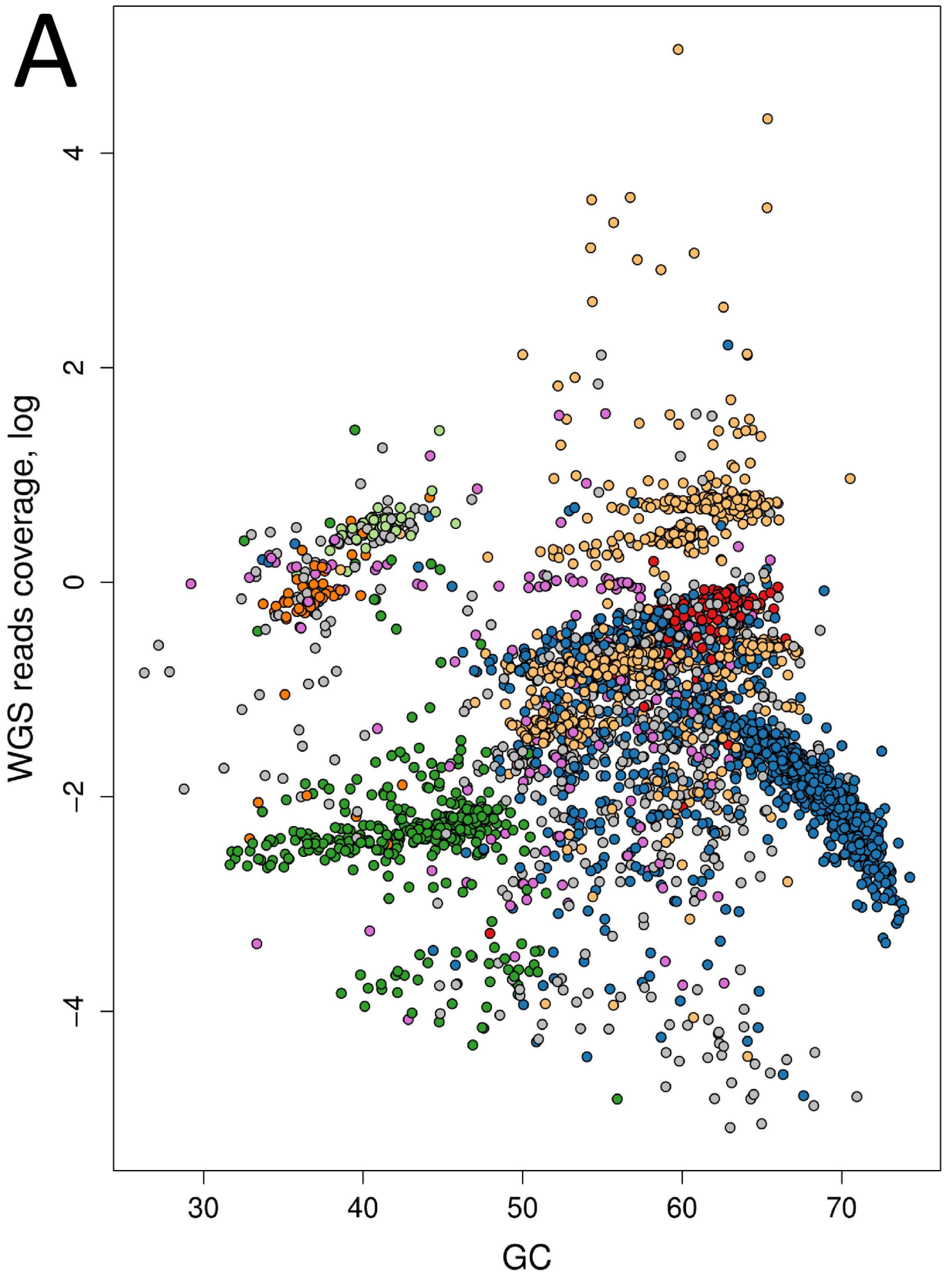
B

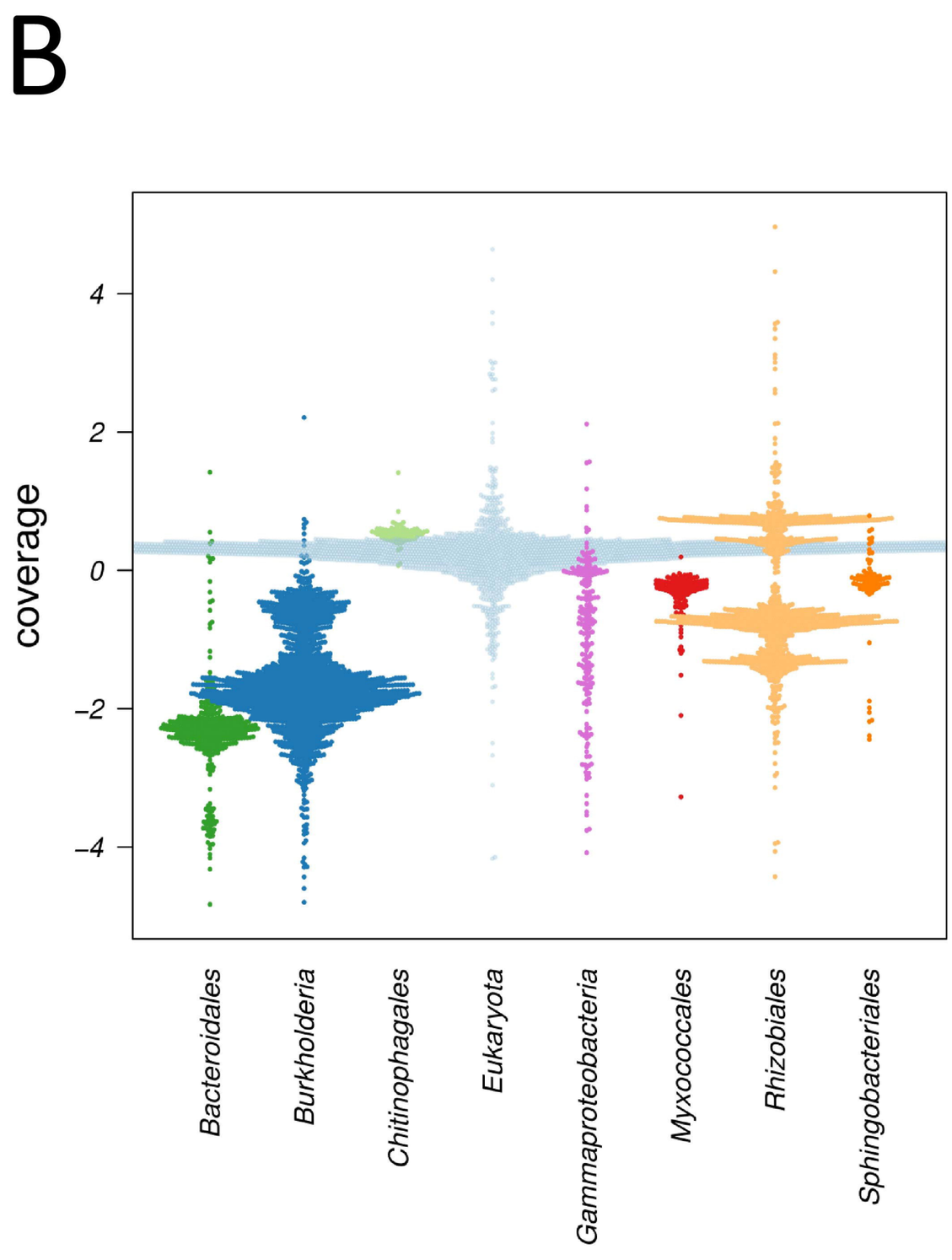


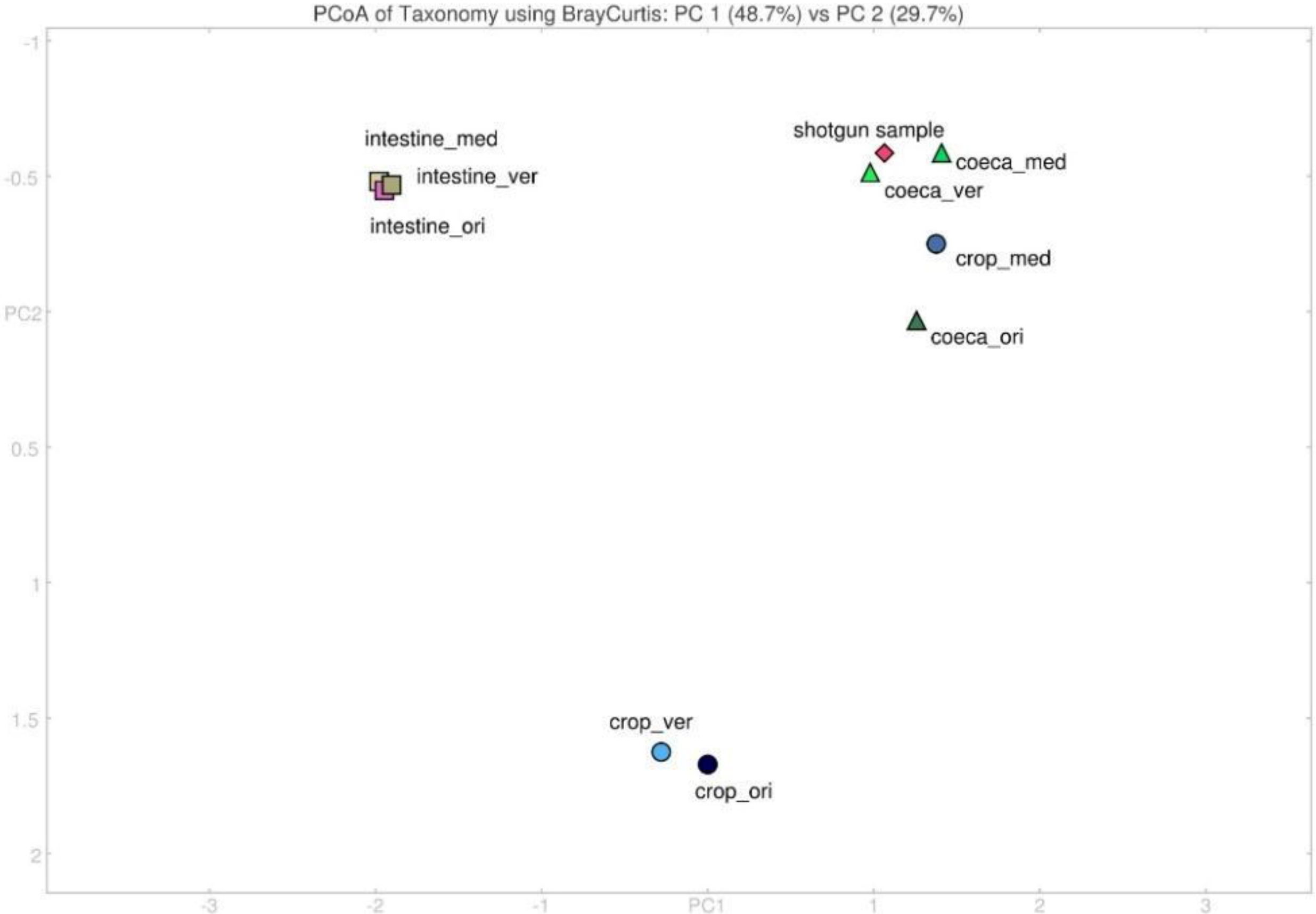
C

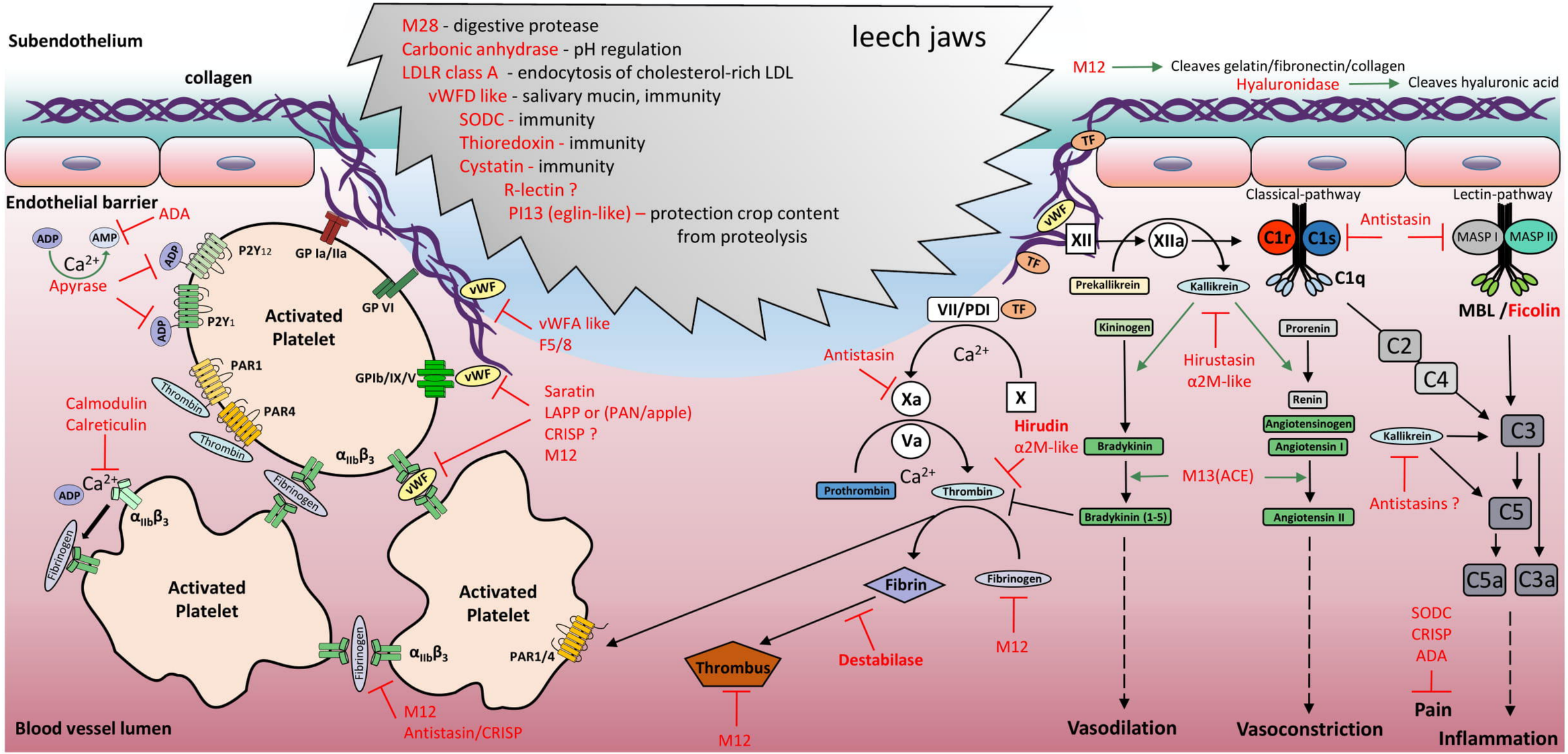



Differential expression orthologous clusters
Proteome orthologous clusters
1
11
13
25
1
2
18
40
40
39
6
1
5
13
H.medicinalis
H.orientalis
H.verbana
80
70
60
50
40
30
20
10
0
DE
DE+Proteom
Proteom
other
NA
identified
Eglin–like
ADA
LDL receptor A
Hyaluronidase
Antistasin
Protease M12
Protease M13
CRISP
NA
Other
Calreticulin
Antistasin
Protease M12
Protease M13
CRISP
Thioredoxin–like
Apyrase
F5/8
Other
NA
SODC
Globin
Tetraspanin
R–type lectin
vWFA–like
Ficolin
Carb_anhydrase
Calmodulin
Protease M28
PAN–domane
a2M
Hyaluronidase
F5/8
vWFD
Antistasin
CRISP
Other
NA


A
WGS reads coverage, log
4
2
0
−2
−4
30
40
50
60
70
GC


B
coverage
4
2
0
−2
−4
Bacteroidales
Burkholderia
Chitinophagales
Eukaryota
Gammaproteobacteria
Myxococcales
Rhizobiales
Sphingobacteriales


PCoA of Taxonomy using BrayCurtis: PC 1 (48.7%) vs PC 2 (29.7%)
intestine_med
intestine_ver
intestine_ori
shotgun sample
coeca_med
coeca_ver
crop_med
coeca_ori
crop_ver
crop_ori
PC2
PC1
-1
-0.5
0.5
1
1.5
2
-3
-2
-1
1
2
3


leech jaws
M28 - digestive protease
Carbonic anhydrase - pH regulation
LDLR class A - endocytosis of cholesterol-rich LDL
vWFD like - salivary mucin, immunity
SODC - immunity
Thioredoxin - immunity
Cystatin - immunity
R-lectin ?
PI13 (eglin-like) – protection crop content from proteolysis
Subendothelium
collagen
Endothelial barrier
M12 → Cleaves gelatin/fibronectin/collagen
Hyaluronidase → Cleaves hyaluronic acid
ADA
ADP
AMP
Ca2+
Apyrase
P2Y12
P2Y1
GP Ia/IIa
GP VI
Activated Platelet
PAR1
PAR4
Thrombin
GPIb/IX/V
vWF
vWFA like
F5/8
Saratin
LAPP or (PAN/apple)
CRISP ?
M12
Calmodulin
Calreticulin
αIIbβ3
Fibrinogen
PAR1/4
M12
Antistasin/CRISP
Blood vessel lumen
TF
VII/PDI
Antistasin
Xa
Va
X
Hirudin
α2M-like
Prothrombin
Thrombin
Fibrin
Fibrinogen
Destabilase
M12
Thrombus
M12
XII
XIIa
Prekallikrein
Kallikrein
Kininogen
Bradykinin
Bradykinin (1-5)
Hirustasin
α2M-like
Prorenin
Renin
Angiotensinogen
Angiotensin I
Angiotensin II
M13(ACE)
Vasodilation
Vasoconstriction
Classical-pathway
Lectin-pathway
C1r
C1s
C1q
Antistasin
MASP I
MASP II
MBL /Ficolin
C2
C4
C3
C5
C5a
C3a
Kallikrein
Antistasins ?
SODC
CRISP
ADA
Pain
Inflammation