Article

Not peer-reviewed version

Revisiting the Asian Buffalo Leech (*Hirudinaria manillensis*) Genome: Focus on Antithrombotic Genes and Their Corresponding Proteins

Zichao Liu , Fang Zhao , Zuhao Huang , Qingmei Hu , Renyuan Meng , Yiquan Lin , Jianxia Qi , Gonghua Lin *

Posted Date: 1 November 2023

doi: 10.20944/preprints202311.0026.v1

Keywords: leech; genome; antithrombotic protein; gene family; hirudin; anticoagulant activity

Article

Revisiting the Asian Buffalo Leech (*Hirudinaria manillensis*) Genome: Focus on Antithrombotic Genes and Their Corresponding Proteins

Zichao Liu [1,†], Fang Zhao [2,†], Zuhao Huang [2], Qingmei Hu [1], Renyuan Meng [1], Yiquan Lin [2], Jianxia Qi [3] and Gonghua Lin [2,*]

[1] Engineering Research Center for Exploitation and Utilization of Leech Resources in Universities of Yunnan Province, School of Agriculture and Life Sciences, Kunming University, Kunming 650214, China
[2] School of Life Sciences, Jinggangshan University, Ji'an 343009, China
[3] Nujiang Management Bureau of Gaoligongshan National Nature Reserve, Nujiang 3036851, China
* Correspondence: lingonghua@163.com
† These authors contributed equally to this work.

Abstract: Leeches are well-known annelids due to their obligate blood-feeding habits. Some leech species secrete various biologically active substances which have important medical and pharmaceutical value in antithrombotic treatments. In this study, we provide a high quality genome of the Asian buffalo leech (*Hirudinaria manillensis*), based on which we performed a systematic identification of potential antithrombotic genes and their corresponding proteins. Combining automatic and manual prediction, we identified 21 antithrombotic gene families including fourteen coagulation inhibitors, three platelet aggregation inhibitors, three fibrinolysis enhancers, and one tissue penetration enhancer. A total of 72 antithrombotic genes, including two pseudogenes, were identified and most of their corresponding proteins forming three or more disulfide bonds. Three protein families (LDTI, antistasin, and granulin) have internal tandem repeats containing 6, 10, and 12 conserved cysteines, respectively. We also measured the anticoagulant activities of the five identified hirudins (hirudin_Hman1 ~ hirudin_Hman5). The results showed that three (hirudin_Hman1, hirudin_Hman2, and hirudin_Hman5), but not the remaining two, exhibited anticoagulant activities. Our study provides the most comprehensive collection of antithrombotic biomacromolecules from a leech to date. These results will greatly facilitate the research and application of leech derivatives for medical and pharmaceutical purposes of thrombosis.

Keywords: leech; genome; antithrombotic protein; gene family; hirudin; anticoagulant activity

1. Introduction

Thrombosis refers to a variety of disorders characterized by aberrant blood clots in the blood that restrict or occlude the lumen of blood vessels, resulting in ischemia and infarction of organs and malfunction [1]. Thrombophilia is the most important component of cardiovascular diseases. According to the World Health Organization, cardiovascular diseases such as ischemic heart disease and stroke directly or indirectly related to thrombophilia claim more than 15 million lives each year worldwide, accounting for 27% of the world's total deaths [2]. Pharmaceutical industries have developed a wide range of anticoagulation, anti-platelet aggregation, and fibrinolysis drugs as thrombosis prevention and therapeutic alternatives. Anticoagulants include warfarin, heparin sodium, argatroban, rivaroxaban; anti-platelet aggregation drugs include aspirin, clopidogrel and abciximab; while fibrinolytic drugs include streptokinase, alteplase and ralteplase [3].

Although antithrombotic drugs have slowed the occurrence of thrombotic events in patients, their success in reducing the lethality of thrombotic disease is limited. The primary reason for this failure is that these drugs generally rely on a single target of action and it is difficult to account for individual differences in dosing, which ultimately leads to frequent drug resistance, internal bleeding, liver and kidney damage, and other serious side effects that put patients' lives at risk. For

 doi:10.20944/preprints202311.0026.v1

instance, warfarin, an anticoagulant, may lead to adverse events including cerebral microhemorrhage, hemorrhagic stroke, subarachnoid hemorrhage, and leukocyte rupture vasculitis [4,5]. The use of the antiplatelet aggregation drug clopidogrel carries a risk of both drug resistance and increased all-cause mortality [6]. On the other hand, the fibrinolytic drugs streptokinase and ralteplase have the potential to cause gastrointestinal bleeding [7]. Thus, the development of safe and effective multi-target drugs with minimal side effects is a crucial direction for the treatment of thrombotic diseases.

Leeches (Hirudinea, Annelida) are well-known annelids due to their obligate blood-feeding habits and are found on every continent except Antarctica [8]. Over 800 species of leeches have been identified [9], and some of them are classified as medicinal leeches [10]. Numerous leeches feed on mammalian blood. To ensure a smooth satiation within a few minutes, they secrete various biologically active substances, including anticoagulants, thrombolytics, and anti-inflammatories, aimed at countering the host's physiological hemostatic process [11]. If used correctly, these bioactive substances can have significant value in medicine and medicinal treatments [12]. The European medical leech (*Hirudo medicinalis*) has been a prevalent tool in the treatment of various ailments in European countries since the 17th century. These conditions include, but are not limited to, osteoarthritis [13], chronic pain [14], cutaneous leishmaniasis [15], and postoperative thrombosis prevention [16,17]. The use of dried leeches, such as *Hirudo nipponia*, is a common practice in traditional Chinese medicine for treating various diseases, particularly those related to thrombosis, including stroke and coronary heart disease. These leeches are often administered orally and have been widely studied by researchers [12,18,19].

There are many antithrombotic active components in leeches [11,20], but the one that has received the most attention is hirudin. Hirudin has a molecular weight of approximately 7000 Daltons and comprises approximately 65 amino acid residues [21,22]. Hirudin is the most potent natural thrombin inhibitor discovered to date, as it binds to thrombin and forms an exceedingly stable non-covalent complex [23,24]. Compared to anticoagulants such as warfarin and heparin, hirudin causes fewer bleeding side effects [25]. In combination with hirudin, leeches possess numerous antithrombotic bioactive compounds within their bodies and secreted saliva. For instance, bdellin, like hirudin, lengthened activated partial thromboplastin time and exerted anticoagulant capability [26]. Destabilase [27], leech carboxypeptidase inhibitor [28], and hyaluronidase [29] were believed to have thrombolytic properties. Saratin and leech antiplatelet protein (LAPP) were reported having anti-platelet aggregation function [30]. Furthermore, antistasin and eglin C have been observed to have analgesic and anti-inflammatory properties and may indirectly affect thrombophilia [31,32]. To our knowledge, over twenty leech-derived antithrombotic active ingredients, involving over fifty gene loci, have been isolated [9,33,34].

There are significant variations in the genomes of different leech species which lead to considerable disparities in the composition, function, and evolution of genes related to thrombosis prevention [35]. Notably, previous studies have shown that the architecture, functional activity, and copy number of genes encoding hirudin can be completely different among diverse leech species [36]. These variations present a challenge to genetic analysis, yet also offer limitless potential for the functional application of these active molecules.

The Asian buffalo leech (*Hirudinaria manillensis* (Lesson, 1842), homotypic synonym: *Poecilobdella manillensis*) is classified under the family Hirudinidae, just like the European medicinal leech *H. medicinalis*. However, it is highly specialized in parasitizing mammals, as opposed to its counterpart. Sawyer [8] noted that this species is commonly found in natural water bodies located in South East Asian countries such as China, Indonesia, Myanmar, the Philippines, Thailand, and Vietnam [37,38]. This species is adapted to feeding on mammalian blood and possesses strong jaws specialized for piercing the tough hides of water buffaloes and elephants [39]. Additionally, the Asian buffalo leech is utilized as a medical and medicinal tool [40]. Adult specimens typically measure 130-140mm in length and 10-13mm in width. In laboratory conditions, a fully-fed Asian buffalo leech can weigh over 50 grams, which is significantly heavier than the average medical leech (*Hirudo* spp.). Furthermore, it has been shown that the antithrombotic properties of the Asian buffalo leech are

 doi:10.20944/preprints202311.0026.v1

superior to those of commonly used *Hirudo* species, including *H. medicinalis* and *H. nipponia* [41]. As a result, there is a great deal of interest in the functional genes of the Asian buffalo leech, and their potential applications in medicine.

The rapid advance of high-throughput sequencing technology means that functional gene research on Asian buffalo leech has reached the level of genomics. Guan et al. [42] reported a draft genome using PacBio RSII sequencing platforms. They created an assembly (GenBank accession: GCA_015345955.1) of 151.8 Mb with a total of 467 scaffold, and an N50 of 2.28 Mb. More recently, Zheng et al. [35] published another genome utilizing the Nanopore PromethION platform. Using chromosome conformation capture (Hi-C) technology, a chromosome-level genome was obtained. It is worth noting that due to the complex structure of the antithrombotic genes, some genes as well as the corresponding proteins were not included in the genome annotation process. Using hirudins as an example, Guan et al. [42] and Zheng et al. [35] identified only one and three hirudins, respectively. However, our investigation revealed the presence of at least five hirudin encoded genes in the two genomes (details below). Given their considerable application potential, high complexity, and intraspecific variability, it is essential to determine the composition and sequence variation of these proteins. Here, we present the chromosome-level genome of an additional sample of the Asian buffalo leech and systematically analyze the proteins related to antithrombotic activities in this species.

2. Materials and Methods

2.1. The DNA and RNA Sequencing

H. manillensis specimens were collected from Fangchenggang, Guangxi, China (coordinates: E107°51′1.47″, N21°41′13.65″). Following removal of their digestive tracts, genomic DNA was collected from fresh tissue using the DNeasy Blood and Tissue Kit (Qiagen, Germany). The quality and integrity of the DNA samples were evaluated through agarose gel electrophoresis, NanoDrop spectrophotometry (NanoDrop Technologies, Wilmington, DE, USA), and Qubit fluorometry (Thermo Fisher Scientific, Waltham, MA, USA). Subsequently, the genomic DNA, which met the standards of quality and quantity, was utilized for building the PacBio and Illumina libraries.

HiFi SMRTbell libraries were prepared with the SMRTbell Express Template Prep Kit 2.0 (PacBio, CA, USA). The genome DNA was fragmented to 15-18 kb using a g-TUBE (Covaris, MA, USA), and damage and fragment ends were repaired with reagents provided in the Template Prep Kit. The repaired ends were then ligated with SMRTbell hairpin adapters, and AMPure PB beads (PacBio, CA, USA) were employed to concentrate and purify the library. To acquire SMRTbell libraries with large inserts, templates larger than 15 kb were selected using the BluePippin system (SageScience, MA, USA). The sequencing procedure was carried out by BioMarker (Beijing, China) utilizing the PacBio Sequel II platform. Thereafter, high-precision HiFi reads were generated using the CCS software (https://github.com/PacificBiosciences/ccs) with default settings.

Hi-C was performed using the following protocol: the leech tissues were fixed in 1% formaldehyde solution. Nuclear chromatin was obtained from the fixed tissue and digested using HindIII (New England Biolabs [NEB], Ipswich, MA, USA). The overhangs were blunted with bio-14-dCTP (Invitrogen, Carlsbad, CA, USA) and Klenow enzyme (NEB). After dilution and religation using T4 DNA ligase (NEB), the genomic DNA was extracted and sheared to 350-500 bp with a Bioruptor (Diagenode, Seraing, Belgium). The biotinlabeled DNA fragments were enriched with streptavidin beads (Invitrogen) and were sequenced using the Illumina HiSeq 2000 sequencing platform (Illumina Inc., San Diego, CA, USA) with both directions of 150 bp reads. We also generated ~100× short reads (called Survey reads) for polishing the genome assembly. DNA libraries with ~350 bp insertions were constructed and were then sequenced with both directions of 150 bp reads using the Illumina HiSeq 2000 sequencing platform. Quality control for raw reads of Hi-C and Survey was performed using fastp 0.20.0 with default settings and parameters [25].

For RNASeq, total RNA was extracted from head tissue using TRIzol reagent (Invitrogen) and purified using the RNeasy Mini Kit (Qiagen, Chatsworth, CA, USA). Oligo(dT)-loaded magnetic

 doi:10.20944/preprints202311.0026.v1

beads were used to purify poly(A) mRNA from total RNA. The mRNA was then fragmented into small pieces (300~500 bp) at 94°C for exactly 5 minutes. The cleaved RNA fragments were reverse transcribed into first-strand cDNA using SuperScript II reverse transcriptase and random primers, and second-strand cDNA was generated using GEX second-strand buffer, DNA polymerase, RNase H and dNTPs. These cDNA fragments were subjected to end repair and 3' adenylation. Paired-end adapters were ligated to the 3′-adenylated cDNA fragments. cDNA fragments of ~300 bp were selected and enriched by 15 cycles of PCR amplification using Phusion DNA Polymerase. Finally, the constructed cDNA libraries were sequenced bidirectionally (150 bp in each direction) using the Illumina HiSeq 2000 sequencing platform. Quality control of raw RNASeq reads was also performed using fastp 0.20.0 with default settings and parameters.

2.2. Genome assembling and gene prediction

The genome was assembled using PacBio HiFi reads through NextDenovo V2.5.0 (https://github.com/Nextomics/NextDenovo). The initial HiFi assembly contigs underwent polishing with Illumina PCR-free data utilizing NextPolish v1.4.0 [43] with recommended settings. For genome assembly at the chromosomal level, Hi-C sequencing reads were utilized with combined usage of samtools v1.3.1 [44], chromap v0.2.5 [45], Juicer v1.6.2 [46], and YaHS programs. The generated files (.hic and .assembly) by YaHS v1.1a were imported into Juicebox v1.11.08 [47] for visualizing Hi-C maps and undergoing manual optimization. Juicer v1.6.2 was employed to generate the final pseudo-chromosome assemblies. In addition, we used the Survey reads to perform mitochondrial genome assembly through GetOrganelle v1.7.7.0 [48] with default settings.

BUSCO (Benchmarking Universal Single-Copy Orthologs) v.4.1.4 [49] with the eukaryota_odb10 database was used to assess the completeness of the genome assembly. We also used Merqury [50] to estimate the quality of the assembly. Both de novo and homology approaches were used to identify repetitive sequences in the leech genomes. First, RepeatModeler v2.0.3 [51] was used to construct the de novo libraries, which were then merged with the Annelida repeat sequences from the RepBase database v20181026 [52]. Finally, RepeatMasker v4.1.2-pl [51] was used to search for repeat sequences, and the repeat-masked genomes were used for gene prediction.

Both ab initio and RNASeq-based prediction strategies were used to predict protein-coding genes in the masked genomes. For the ab initio prediction strategy, GlimmerHMM v3.0.4 [53] and SNAP v2006-07-28 [54] were used with default settings to predict genes in the genome sequences. As for the RNASeq-based prediction strategy, three approaches were applied: (1) PASA approach, Trinity v2.9.0 [55] was used to assemble RNASeq reads to generate unigenes, and gene information was predicted using PASA v 2.5.2 [56]. (2) Stringtie approach, HISAT v2.147 [57] was used to align RNASeq reads to the genome, and gene information was predicted using StringTie v1.3.4c48 [58]. (3) BRAKER approach, STAR v2.7.9a [59] was used to align the RNASeq reads to the genome, and gene information was predicted using BRAKER v2.1.6 [60]. Finally, the General Feature Format (GFF) files generated from the above approaches were combined and integrated using EVidenceModeler v1.1.1 [61] to generate high-confidence gene sets (evm.gff). The coding sequences (CDS) and the corresponding protein sequences of all protein-coding genes were extracted using gffread v0.12.7 [62]. All protein sequences were functionally annotated by the NCBI NR, Uniprot TrEMBL, eggNOG [63] and Pfam [64] databases using BlastP with an E-value ≤1e-5.

2.3. Identification of antithrombotic genes

To the best of our knowledge, we collected all available literatures on leech antithrombotic genes and/or proteins. Sequence information was obtained directly from the literature or retrieved from genetic databases (GenBank, Uniprot, etc.) according to the accession number in the literature. All coding sequences were translated into protein sequences using MEGA software. These reported leech antithrombotic proteins (RAPs) were used as query sequences to identify homologous proteins in the *H. manillensis* genome. It should be mentioned that the above automated gene prediction approaches varied in their predictive sensitivity to antithrombotic genes. Even the relatively best BRAKER approach could predict only three out of five hirudins (see below). Fortunately, there are many

 doi:10.20944/preprints202311.0026.v1

RNASeq reads of *H. manillensis* (accession SRR15881191~ SRR15881251) in the sequence reads archive of GenBank, which greatly facilitate our manual prediction on antithrombotic genes.

To obtain as many antithrombotic genes as possible, we used a so-called BRAKER-plus strategy to improve the prediction of potential antithrombotic genes. First, we assembled the RNA-Seq data sequenced in this study and those from GenBank using Trinity. The generated unigenes were used for CDS prediction using GeneMarkS-T v5.1 [65]. Second, we used the RAPs as queries to blast and extract all candidate antithrombotic genes from the RNASeq-derived CDS. All of these candidate genes were then mapped to the *H. manillensis* genome using the est2genome model in Exonerate v2.2.0 [66]. The highly matched regions were retained and manually checked to remove redundant or low quality information. Third, the GFF files from the manual prediction and the above BRAKER prediction were merged using AGAT v1.2.0 [67]. After cleaning duplicated features between manual prediction and BRAKER prediction, we obtained a final version of the GFF file called BRAKER-plus.gff, which had updated information on antithrombotic genes.

2.4. Variation analysis of antithrombotic proteins

The CDS of all antithrombotic genes were translated into protein sequences using Seqkit v0.10.2 [68]. The signal peptide region of each protein was predicted using the online tool SinalP (https://dtu.biolib.com/SignalP-6). For each gene family, the proteins predicted in this study and the corresponding archetypal protein were combined and aligned using Clustal W in MEGA v11.0.13 [69]. The protein families with extremely high mutation rates were manually verified. The pairwise sequence similarity of the proteins was calculated using the water program from EMBOSS v 6.6.0.0 [70]. The longest similarity index, (similar residues with BLOSUM62 matrix) × 100) / (length of the alignment - total number of gaps in the alignment), was used.

The relationships among members of the hirustasin superfamily and the HMEI family are too complex to be directly observed. We reconstructed phylogenetic trees of the two protein families using IQ-TREE v1.6.12 [71]. The best-fitting models were tested using the ModelFinder module embedded in IQ-TREE. As suggested by the authors of this program, 1000 bootstrap replicates were set to test the support rates of the tree nodes. In addition, to test the potential relationships between anticoagulant activity and molecular evolution of hirudins, we also collected the sequence and functional information of hirudins from previous reports. We then combined the reported hirudins and those identified in this study and reconstructed the phylogenetic relationships of all sequences using IQ-TREE.

2.5. Anticoagulation analyses of hirudins

Hirudin was the first identified and most studied antithrombotic protein. The antithrombotic (mainly anticoagulant) activities of hirudin variants from different leech species have been repeatedly analyzed [22]. We used *Pichia pastoris* as expression vector to produce recombinant hirudin variants of *H. manillensis* according to Hu et al. [72] with minor modifications. Briefly, the DNA sequences of hirudin variants were synthesized by chemical synthesis methods and subcloned into pPIC9K-His using EcoRI and NotI restriction enzymes to obtain recombinant plasmids. The recombinant plasmids were linearized with SalI and electroporated into *P. pastoris* strain GS115. The cells were plated on yeast extract peptone-dextrose medium containing 0, 0.25, 0.50, 1.0 and 2.0 mg/ml G418. The positive transformants harboring recombinant hirudin variants were identified by PCR and grown overnight in 5 ml buffered glycerol-complex medium at 30°C with shaking. After 2 days of culture, hirudin protein expression was induced by daily addition of 1.0% v/v methanol. The harvested supernatant was then centrifuged and the protein of interest was extracted with 70% ammonium sulfate.

The antithrombin activity of the recombinant protein was evaluated using the antithrombin titration method as previously described [73], with minor modifications. First, three reagents were prepared: A, 0.5 mg/mL harvested hirudin solution diluted in 0.2mol/L PBS (phosphate buffer solution, pH = 7.4); B, 10 mg/mL bovine fibrinogen solution in normal saline; C, thrombin solution with 50 U/mL anticoagulant activity in normal saline. Second, to prepare a 2 ml glass tube, add 100

 doi:10.20944/preprints202311.0026.v1

μL of hirudin solution and 200 μL of reagent B to the tube, and then place the tube in a 37℃ water bath for 5 minutes. Third, add 5 μL of Reagent C to the tube every 4 minutes while gently shaking. Fourth, stop adding Reagent C when the liquid starts to solidify and calculate the anticoagulation activity U = (C1 * V1) / (C2 * V2). C1, V1, C2, and V2 indicate the concentration of Reagent C, the total volume of Reagent C added to the tube, the concentration of Reagent A, and the volume of Reagent A used in the reaction, respectively. PBS solution without any proteins was used as control group. For each group, three duplicates were set.

3. Results

3.1. This Genome sequencing and assembling

We used the PacBio HiFi platform for TGS sequencing and performed genome assembly for *H. manillensis*. A total of 23.80 Gb of highly accurate HiFi reads were obtained, with an average length of 18.73 Kb. The de novo assembly generated 34 contigs with a total length of 149.59 Mb and an N50 of 9.58 Mb. We then used Illumina HiSeq platform for NGS sequencing of Hi-C libraries, and based on the obtained 21.09 Gb Hi-C reads, we consolidated the contigs into 22 scaffolds with total length and N50 of 148.59 Mb and 11.46 Mb, respectively. The first 13 longest scaffolds were 14.76 ~ 8.30 Mb in length, while the remaining 9 scaffolds were less than 0.3 Mb each. The highly discontinuous scaffold length distribution and the well-resolved Hi-C maps (Figure 1) clearly represent 13 chromosomes of this species. The 13 chromosomes had a total length of 147.43 Mb, representing 99.22% of the total scaffold length. In addition, we used the Illumina HiSeq platform for NGS sequencing and performed mitochondrial genome assembly using GetOrganelle. A total of 9.52 Gb NGS reads were sequenced and a circular complete mitochondrial genome of 15,581 bp was obtained. As a result, we obtained a nearly complete genome of *H. manillensis* with a total length of ~150 Mb, including 13 chromosomes, one mitochondrial genome, and 9 debris (1.16 Mb).

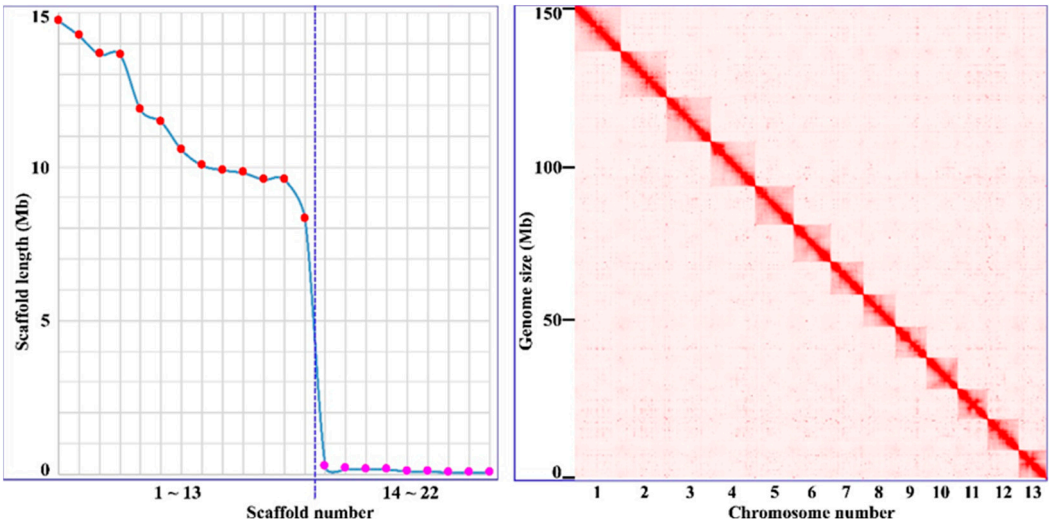

Figure 1. Scaffold and chromosome lengths of *H. manillensis* genome.

We estimated the completeness of the final assemblies using BUSCO (Benchmarking Universal Single-Copy Orthologs) with the eukaryota_odb10 database. The results showed that 98.0% of the BUSCOs were captured, including 92.5% complete and single-copy BUSCOs, 3.1% complete and duplicated BUSCOs, and 2.4% fragmented BUSCOs. We also used Merqury to estimate the quality of the assembly and obtained a quality score of 33.53. The BUSCO and Merqury results indicated that we had obtained a high completeness and high quality genome of this species. We also searched for repeat sequences in the genomes using RepeatModeler and RepeatMasker. A total of 18.75% of the genome was identified as repeats, including retroelements (6.59%), DNA transposons (1.76%), rolling circles (0.20%), unclassified elements (8.41%), satellites (0.2%), and simple repeats (1.8%). No small

 doi:10.20944/preprints202311.0026.v1

RNA or low complexity repeat elements were found. The masked genome was used for gene prediction.

3.2. Gene prediction and annotation

We used two ab initio prediction approaches (GlimmerHMM and SNAP) and three RNASeq-based prediction approaches (PASA, Stringtie, and BRAKER) to predict protein-coding genes in the masked genomes. The ab initio approaches detected more genes than the RNASeq-based approaches, but with much lower N50 values of the CDS (< 1 kb). Of the three RNASeq-based prediction approaches, BRAKER detected the smallest number of genes (25,331), but had the largest N50 values of the CDS (2,151 bp) (Table 1). After combination by EVidenceModeler, a total of 21,828 genes were predicted with a CDS N50 of 1,731 bp.

Table 1. Statistics of predicted protein-coding genes in *H. manillensis* genome.

Annotation approach	Gene number	Total CDS length (Mb)	N50 of CDS (bp)	No. of hirudins
GlimmerHMM	56,061	31,047,090	873	2
SNAP	45,826	26,341,403	762	2
PASA	37,111	33,145,509	1,122	0
Stringtie	32,919	40,238,799	1,554	0
BRAKER	25,331	36,676,815	2,151	3
EVidenceModeler	21,828	26,863,541	1,731	0

To test the prediction sensitivity, we used the well-known hirudin (HV1) as a query to blast the CDS predicted by each approach. The results showed that two hirudin-coding genes were predicted by each of the two ab initio prediction approaches, and three were predicted by the BRAKER approach. Unexpectedly, no such gene was detected by PASA, Stringtie, or the EVidenceModeler integrated sequences (Table 1). Considering all the above factors, we decided to use the CDS and proteins predicted by BRAKER as the relatively best prediction results for further analyses.

We aligned the protein sequences predicted by the BRAKER approach against the NCBI NR, Uniprot TrEMBL, eggNOG, and Pfam databases to perform functional annotations. About 2/3 of the proteins were functionally assigned to NR, TrEmbl, EggNOG, and Pfam databases, respectively (Table 2). After synthesizing the results of the four assignments, a total of 18,373 (72.53%) genes were functionally assigned or annotated.

Table 2. Functional annotation of proteins predicted from the *H. manillensis* genome.

Annotation approach	Gene number	Percentage
NR	17,411	68.73%
TrEmbl	17,421	68.77%
EggNOG	15,952	62.97%
Pfam	15,675	61.88%
integration	18,373	72.53%

3.3. Identification of antithrombotic genes

By reviewing the literature on leeches, we collected 21 reliable antithrombotic proteins, including 14 coagulation inhibitors, three platelet aggregation inhibitors, three fibrinolysis enhancers, and one tissue penetration enhancer. The anticoagulants can also be further categorized into three groups: (1) thrombin inhibitors, including hirudin and progranulin; (2) Factor Xa inhibitors, including antistasin, lefaxin, and therostasin; and (3) serine protease inhibitors, including hirustasin, guamerin, piguamerin, bdellastasin, poecistasin, eglin, bdellin, leech-derived tryptase inhibitor (LDTI), and *H. manillensis* elastase inhibitor (HMEI) (Table 3). We used the archetypal antithrombotic proteins as queries to blast the gene set predicted by the BRAKER approach. A total of 61

 doi:10.20944/preprints202311.0026.v1

antithrombotic genes were identified, and at least one homologous gene was detected for most of the gene families (19 out of 21 gene families, except for therostasin and bdellin). Similarly, we blasted the predicted genes from the previous studies on *H. manillensis* [35,42]. A total of 28 antithrombotic genes from 12 gene families were detected by Guan et al. [42], while 46 antithrombotic genes from 17 gene families were detected by Zheng et al. [35] (Table 4).

Table 3. Archetypal sequence of antithrombotic proteins from leeches (note: as an exception, *Lumbricus rubellus* is an earthworm).

No.	Protein	Leech species	Accession/reference	Function
1	hirudin	*H. medicinalis*	ALA22933.1	coagulation inhibitor
2	granulin	*H. nipponia*	[83]	coagulation inhibitor
3	antistasin	*Haementeria officinalis*	AAA29192.1	coagulation inhibitor
4	lefaxin	*Haementeria depressa*	P86681.1	coagulation inhibitor
5	therostasin	*Theromyzon tessulatum*	AAF73958.1	coagulation inhibitor
6	hirustasin	*H. medicinalis*	P80302.1	coagulation inhibitor
7	guamerin	*H. nipponia*	P46443.1	coagulation inhibitor
8	piguamerin	*H. nipponia*	P81499.1	coagulation inhibitor
9	bdellastasin	*H. medicinalis*	P82107.1	coagulation inhibitor
10	poecistasin	*H. manillensis*	[96]	coagulation inhibitor
11	eglin	*H. medicinails*	PDB: 4H4F	coagulation inhibitor
12	bdellin	*H. nipponia*	AAK58688.1	coagulation inhibitor
13	LDTI	*H. medicinails*	P80424.1	coagulation inhibitor
14	HMEI	*H. manillensis*	[107]	coagulation inhibitor
15	saratin	*H. officinalis*	PDB: 2K13	platelet aggregation inhibitor
16	apyrase	*Helobdella robusta*	XP_009028854.1	platelet aggregation inhibitor
17	lumbrokinase	*L. rubellus*	AAN28692.1	platelet aggregation inhibitor
18	destabilase	*H. medicinails*	AAA96143.1	fibrinolysis enhancer
19	GGT	*H. medicinails*	[124]	fibrinolysis enhancer
20	LCI	*H. medicinails*	[28]	fibrinolysis enhancer
21	hyaluronidase	*H. nipponia*	AHV78514.1	tissue penetration enhancer

Table 4. Identified antithrombotic genes from *H. manillensis* genome in previous studies and in this study.

No.	Gene family	Guan et al. [42]	Zheng et al. [35]	BRAKER prediction	BRAKER-plus prediction
1	*hirudin*	1	3	3	5
2	*progranulin*	1	1	1	1
3	*antistasin*	0	2	2	2
4	*lefaxin*	3	3	3	3
5	*therostasin*	0	1	0	1
6	*hirustasin / hirustasin-like*	1 / 3	1 / 5	1/11	1 / 12
7	*guamerin*	0	0	1	1
8	*piguamerin*	0	0	1	1
9	*bdellastasin*	0	1	1	1
10	*poecistasin*	0	1	0	2
11	*eglin*	0	3	3	4
12	*bdellin*	0	0	0	1
13	*LDTI*	1	1	1	1
14	*HMEI*	4	9	15	18
15	*saratin*	1	1	2	2

 doi:10.20944/preprints202311.0026.v1

16	*apyrase*	5	5	5	5
17	*lumbrokinase*	3	3	3	3
18	*destabilase*	0	1	3	3
19	*GGT*	1	1	1	1
20	*LCI*	1	0	1	1
21	*hyaluronidase*	3	3	3	3
—	total	28	46	61	72

To obtain as many antithrombotic genes as possible, we used a so-called BRAKER-plus strategy, which combined manual prediction and BRAKER prediction to improve the prediction of antithrombotic genes. Based on the BRAKER-plus strategy, a total of 72 antithrombotic genes (70 integrated and two pseudogenetic) were successfully identified from the 21 gene families, outperforming previous studies. Of the 72 corresponding antithrombotic proteins, 59 were encoded by 12 multi-gene families (hirudin, antistasin, lefaxin, bdellastasin, poecistasin, eglin, HMEI, saratin, apyrase, lumbrokinase, destabilase, and hyaluronidase) that contained two or more closely related members. The remaining nine proteins (progranulin, therostasin, hirustasin, guamerin, piguamerin, bdellin, LDTI, GGT, and LCI) were encoded by single gene families. Nine of the 12 multi-gene families were clustered in a single chromosome, while three families (HMEI, lumbrokinase, and hyaluronidase) were involved in two or three chromosomes.

3.4. Genetic variation of antithrombotic proteins

Hirudin is the first antithrombotic bioactive molecule identified in leeches [74] and has been used for many years as an anticoagulant because it binds directly to thrombin to prevent blood clotting [75]. The first complete amino acid [76] and cDNA [77] sequences of hirudin have been determined from *H. medicinalis*. Subsequently, additional subtypes or variants have been sporadically reported from *H. medicinalis* [78], *H. manillensis* [79], *Haemadipsa sylvestris* [80], *H. nipponia* [26], *Whitmania pigra* [81]. However, there has been no genome wide systematic study of hirudins and their encoding genes in a leech species. Here, five hirudin genes were found in the *H. manillensis* genome. The corresponding five proteins of these genes were 66 ~ 85 amino acid residues in length. Online blasting showed that these proteins had 57.63% ~ 77.83% identity with the hirudins deposited in GenBank (Table 5), indicating that these proteins were all hirudin variants. There was a relatively conserved signal peptide region and six cysteine sites (except for hirudin_Hman3 which had a C34R mutation) (Figure 2), which was a typical nature of hirudins to form three disulfide bonds [80]. The sequence similarities between the archetypal hirudin (HV1, ALA22933.1) and those identified in this study were 62.96% ~ 78.95%.

Table 5. NCBI online blast results of the five hirudins predicted in this study.

hirudin	Target species	Protein	Accession	Identity
hirudin_Hman1	*H. manillensis*	HM1	Q07558.1	76.83%
hirudin_Hman2	*H. manillensis*	HLF8	APA20852.1	57.63%
hirudin_Hman3	*H. manillensis*	HLF7	APA20868.1	62.96%
hirudin_Hman4	*H. manillensis*	HLF6	APA20866.1	73.68%
hirudin_Hman5	*H. manillensis*	HM1	Q07558.1	68.33%

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

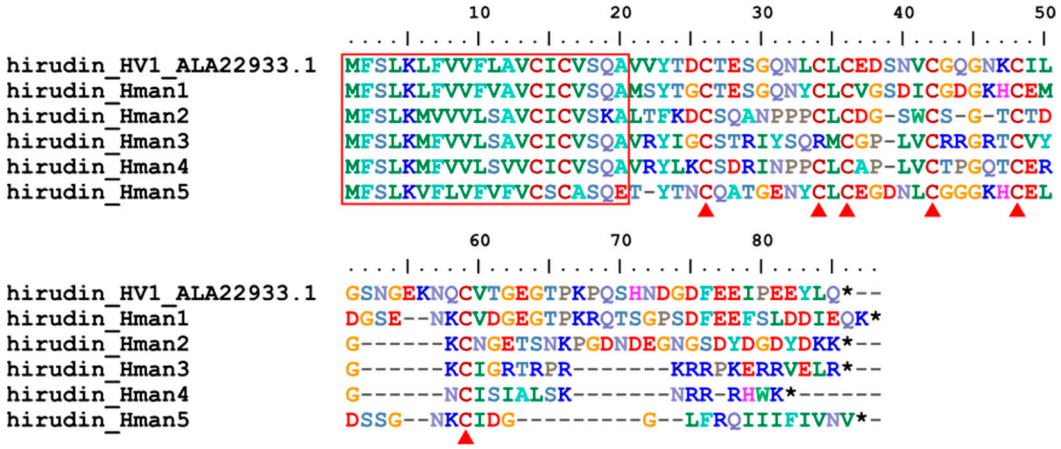

Figure 2. Alignment of archetypal hirudin (from *H. medicinalis*) and those from this study.

Granulin is a secreted protein that acts as a key regulator of lysosomal function and as a growth factor involved in inflammation, wound healing, and cell proliferation. The precursor protein, progranulin, consists of a highly conserved tandem repeat of 12 cysteines in the primary sequence [82]. The leech granulin isolated from *H. nipponia* behaved as a thrombin inhibitor [83], although the mechanism was not clear. Here, a single progranulin gene was detected in the *H. manillensis* genome. The corresponding precursor protein contained 478 amino acids including a signal peptide region. The archetypal granulin peptide from *H. nipponia*, which was first shown to have anti-thrombin activities [83], is located in the N-terminus of the proteins. Five conserved internal tandem repeats were found (Figure 3), each containing 12 conserved cysteines. The sequence similarity between the archetypal granulin and *H. manillensis* progranulin was 95.45%.

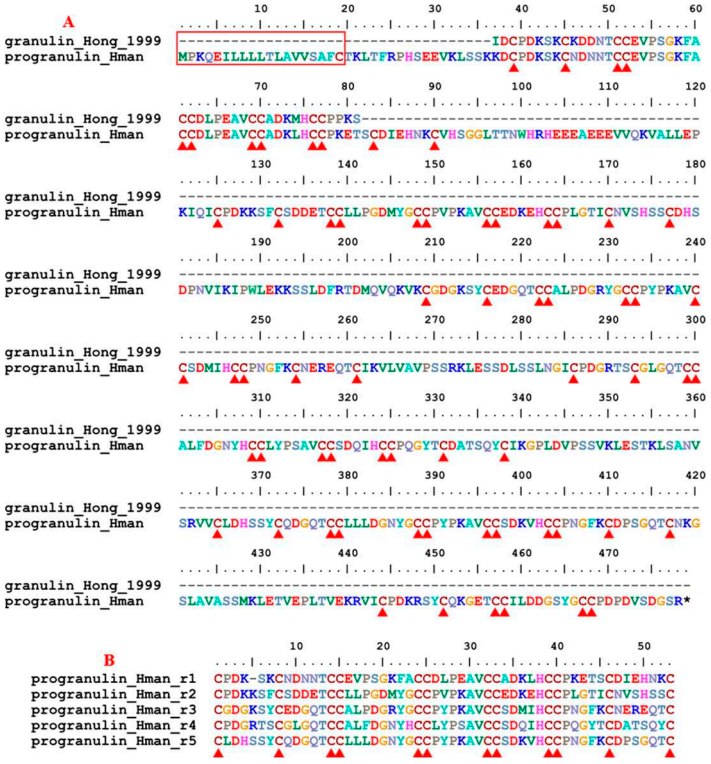

Figure 3. (A) Alignment of archetypal granulin (from *H. nipponia*) and that from this study; (B) Alignment of the five internal tandem repeats of *H. manillensis* progranulin.

Antistasin belongs to a class of serine protease inhibitors characterized by a well-conserved pattern of cysteine residues [84]. Antistasin is a potent anticoagulant that stoichiometrically and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

selectively inhibits mammalian Factor Xa, a pivotal serine protease in the coagulation cascade [85]. Antistasin has been identified in two *Haementeria* species, *H. officinalis* [86] and *H. ghilianii* (nominated as *ghilanten*) [87], but no homology has been found in other leeches. Here, two antistasin genes were detected in the genome and the corresponding protein sequences were relatively conserved. Similar to the archetypal antistasin from *Haementeria* [86,88], there were 20 cysteine residues (Figure 4) in all antistasin proteins. In addition, two internal tandem repeats were observed in each of the two antistasin proteins (Figure 4), confirming that these molecules are true antistasins. The sequence similarities between the archetypal antistasin (AAA29192.1) and those identified in this study were 58.33% ~ 61.46%.

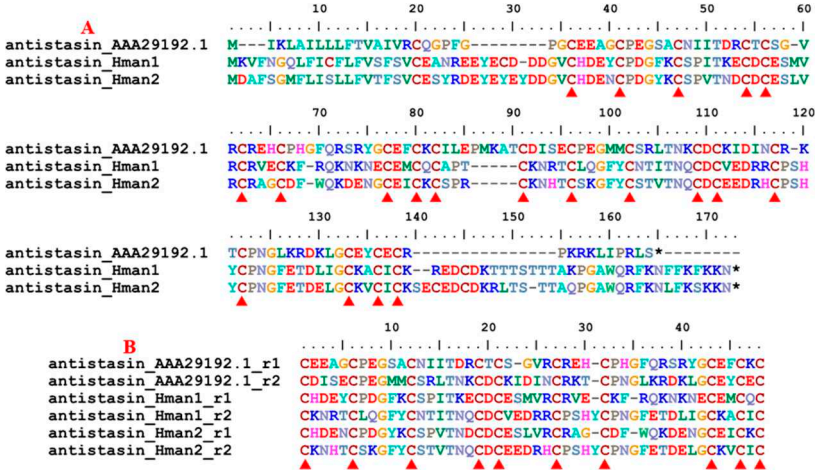

Figure 4. (A) Alignment of archetypal antistasin (from *H. officinalis*) and those from this study; (B) Alignment of the five internal tandem repeats of *H. manillensis* antistasin.

Lefaxin (leech Factor Xa inhibitor), as its name suggests, was also an effective inhibitor of Factor Xa. Lefaxin was first purified from dissected salivary complexes of a leech *H. depressa*. Inhibition of Factor Xa by lefaxin was demonstrated by inhibition of its amidolytic activity and its ability to inhibit thrombin generation in the prothrombinase complex [89]. Three genes have been identified in the *H. manillensis* genome. In contrast to the other Factor Xa inhibitors, no signal peptide or conserved cysteines were found in the corresponding proteins. The archetypal lefaxin from *H. depressa* corresponded to the N-terminus of the six lefaxins (Figure 5). The sequence similarities between the archetypal lefaxin (P86681.1) and those identified in this study were 75.00% ~ 78.57%.

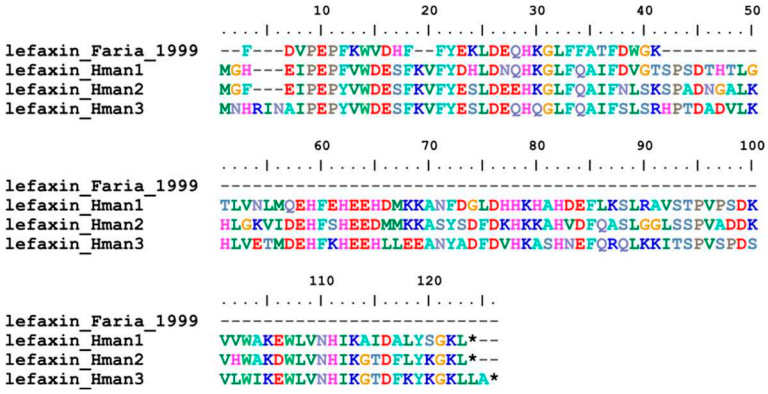

Figure 5. Alignment of archetypal lefaxin (from *H. depressa*) and those from this study.

Therostasin is another potent naturally occurring inhibitor of mammalian Factor Xa. It was first isolated from the duck leech *T. tessulatum*. Like the other known inhibitors of Factor Xa, Therostasin is expressed and stored in the cells of the salivary glands of the leech. Therostasin is a cysteine-rich protein with 16 cysteine residues [90]. A therostasin gene has been identified in *H. manillensis*. The corresponding protein sequences contain a signal peptide region that is much longer than that of *T.*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

tessulatum. In contrast, the functional region was much more conserved between the two leeches, especially the 16 cysteine residues (Figure 6). The sequence similarity between the archetypal therostasin (AAF73958.1) and that of *H. manillensis* was 72.45%.

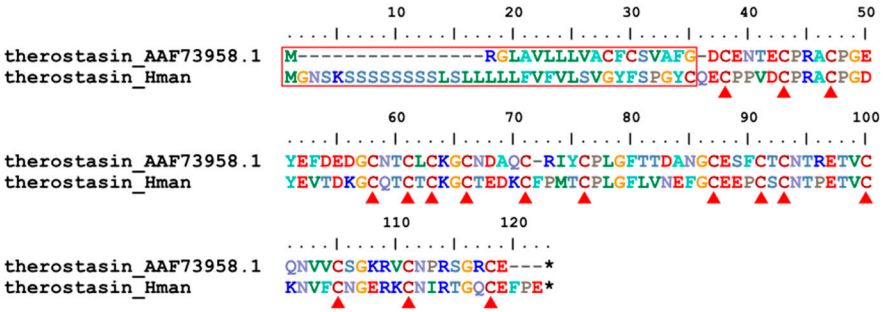

Figure 6. Alignment of archetypal therostasin (from *T. tessulatum*) and that from this study.

Hirustasin was a serine protease inhibitor first identified from *H. medicinalis* [91]. Unlike antistasin or therostasin, hirustasin neither inhibited in vitro blood coagulation nor the amidolytic activity of isolated mammalian Factor Xa. In contrast, hirustasin inhibited trypsin, chymotrypsin, tissue kallikrein, and neutrophil cathepsin G [91], which have activating effects on Factor X, and enhanced activity on Factor XII [11]. Later, additional analogues of hirustasin were identified from other leeches: guamerin and piguamerin from *H. nipponia* [92,93], guamerin variant II from *Whitmania edentula* [94], bdellastasin from *H. medicinalis* [95], and poecistasin from *H. manillensis* [96]. Due to the high similarity in both sequence and function, we referred to these proteins as hirustasin superfamily members. In this study, a total of 18 hirustasin superfamily members were detected. We combined all these members and the archetypal hirustasin (P80302.1), guamerin (P46443. 1) as well as guamerin II [94], piguamerin (P81499.1), bdellastasin (P82107.1), and poecistasin [96], and after alignment, we tried to determine their affiliation by phylogenetic analysis using IQ-TREE. The results showed that one hirustasin, one guamerin, one piguamerin, one bdellastasin, and two poecistasin were identified. The remaining 12 proteins were not clustered to hirustasin and its analogues, hence were nominated as hirustasin-like proteins for convenience. All proteins had a signal peptide and 10 conserved cysteines (Figure 7). The sequence similarities between the archetypal hirustasin (P80302.1) and those identified in the *H. manillensis* genome were 55.32% ~ 78.72%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

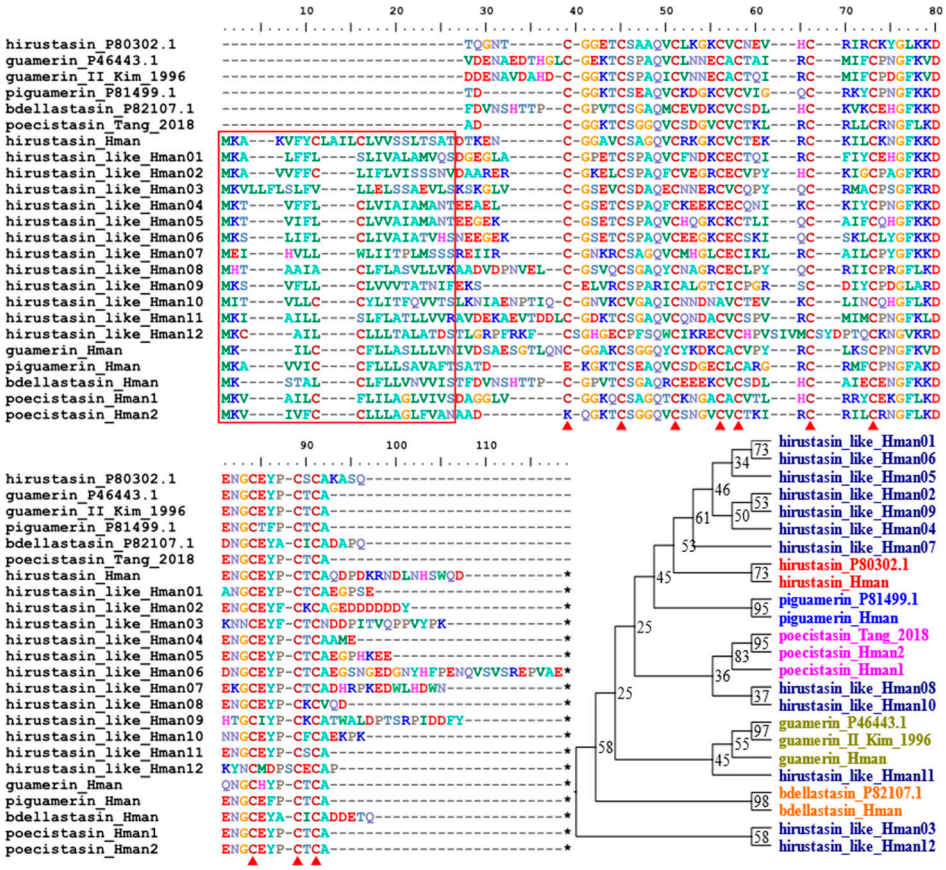

Figure 7. (A) Alignment and (B) phylogenetic relationships of archetypal hirustasin (*H. medicinalis*), guamerin (*H. nipponia*) and piguamerin (*H. nipponia*), bdellastasin (*H. medicinalis*), poecistasin (*H. manillensis*) and their homologous proteins identified in this study.

Eglin (elastase-cathepsin G leech inhibitor) is a small protein with potent inhibitory activity against chymotrypsin and subtilisin-like serine proteinases acting on non-cationic substrates [97]. This molecule was first identified from *H. medicinails* and is able to inhibit platelet activation induced by cathepsin G [98]. Two eglin variants (eglin B and eglin C) were detected from a single species [99], suggesting that they are encoded by a multi-gene family. Four eglin genes were identified in the genome of *H. manillensis*, and the corresponding proteins were relatively conserved (Figure 8). The signal peptide region was detected, but as previously reported [99], no cysteine residues were found in the proteins. The sequence similarities between the archetypal eglin (PDB accession: 4H4F) and those identified in this study were 83.08% ~ 88.41%.

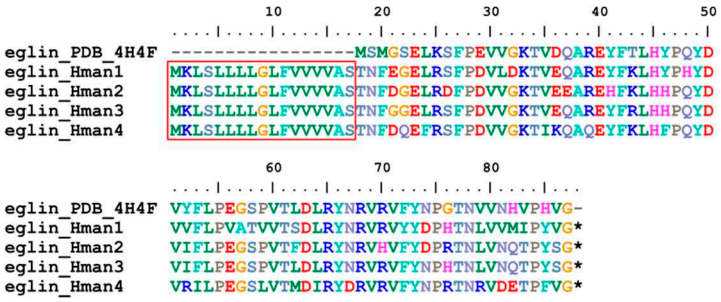

Figure 8. Alignment of archetypal eglin (from *H. medicinalis*) and those from this study.

Bdellin is an inhibitor of trypsin, plasmin and sperm acrosin. It was first discovered in 1969 [100] and its protein sequence was first identified from *H. medicinalis* [101]. It has Kazal serine proteinase inhibitors with a typical domain of six cysteine residues forming a 1-5, 2-4, 3-6 disulfide bond pattern [102]. A bdellin gene was detected in the *H. manillensis* genome. The corresponding protein sequences

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

were relatively conserved, including a signal peptide region of 19 amino acids, an N-terminal region of about 40 amino acids and six cysteines, and a C-terminal region with a highly charged repetitive sequence (Figure 9). The sequence similarity between the archetypal bdellin (AAK58688.1) and that of *H. manillensis* was 78.17%.

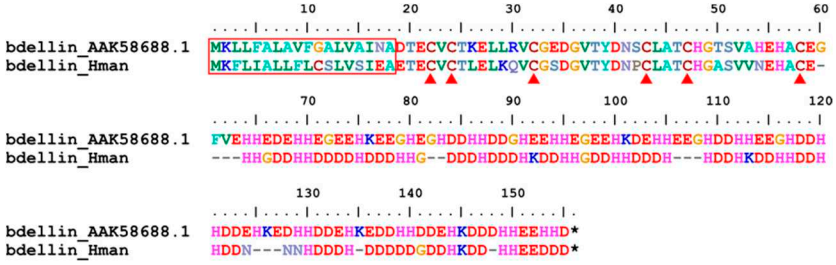

Figure 9. Alignment of archetypal bdellin (from *H. medicinalis*) and that from this study.

Leech-derived tryptase inhibitor (LDTI) is another Kazal-type serine protease inhibitor that inhibits mast cell tryptase in particular, but also trypsin and chymotrypsin [103]. LDTI was first isolated from *H. medicinalis* [104]. Recombinant LDTI was shown to inhibit thrombin and trypsin, thereby prolonging blood clotting time [105,106]. Here, we identified a single LDTI gene in the *H. manillensis* genome. There was a signal peptide region in the corresponding proteins. The rest of the functional region was also relatively conserved, especially on the 12 cysteines (Figure 10). Sequence alignment showed that the archetypal LDTI from *H. medicinalis* matched the N-terminal of the functional regions of the LDTI from *H. manillensis*. Further analysis revealed that, similar to antistasin mentioned above, two internal tandem repeats of six cysteines are observed in these proteins (Figure 10). The sequence similarity between the archetypal LDTI (P80424.1) and that from *H. manillensis* was 92.50%.

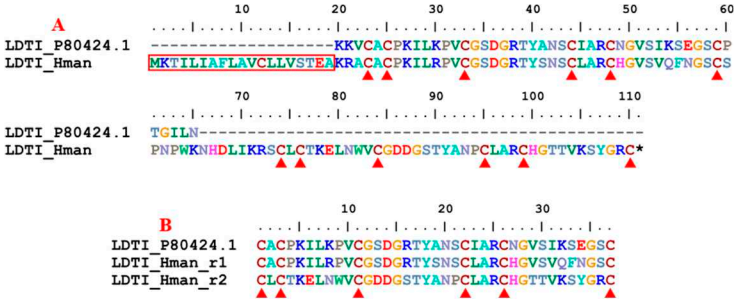

Figure 10. (A) Alignment of archetypal LDTI (from *H. medicinalis*) and that from this study; (B) Alignment of the archetypal LDTI and two internal tandem repeats of *H. manillensis* LDTI.

HMEI (*H. manillensis* elastase inhibitor) is a newly discovered elastase inhibitor from *H. manillensis*. The protein belongs to a multigene family because at least two members, HMEI-A and HMEI-B, have been detected. The HMEI-A showed potent abilities to inhibit elastase and as a result inhibited the formation of neutrophil extracellular trap (NET). Although the authors did not test the activity of HMEI-B, a similar function is expected based on the high sequence identity, especially the highly conserved cysteine residues [107]. Since NET plays an important role in abnormal thrombus formation [108], its inhibition by HMEI proteins will lead to additional antithrombus outcomes. A total of 18 HMEI genes were detected in the *H. manillensis* genome, indicating that they belong to a large gene family. Of the 18 corresponding proteins three (HMEI_Hman03 ~ HMEI_Hman05) were identical to each other. Phylogenetic analysis showed that HMEI_Hman16 and HMEI_Hman15 were closely related to HMEI-A and HMEI-B, respectively (Figure 11). The proteins were highly variable in both sequence length and amino acid sites. The proteins had a signal peptide and ten cysteines were clearly visible. The pairwise similarity between the archetypal HMEI-A [107] and the newly detected proteins was 52.83% ~ 99.41%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

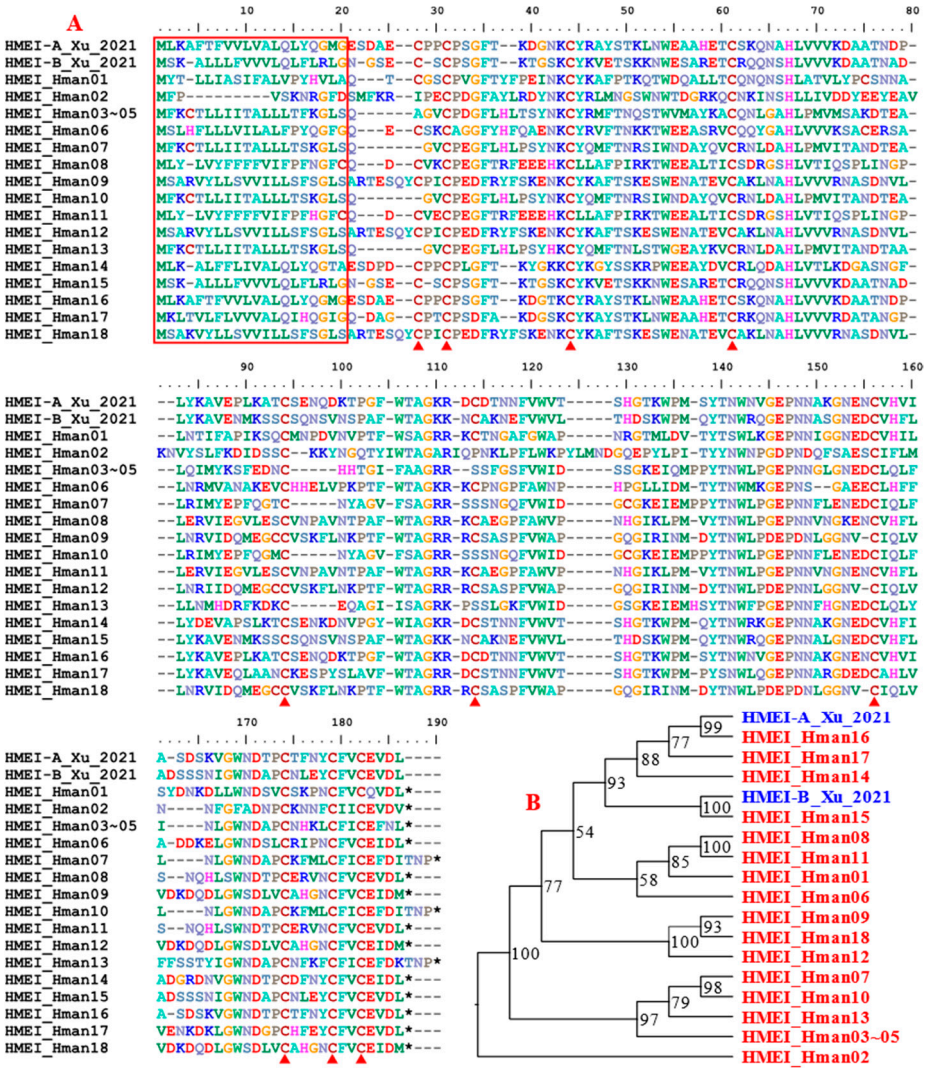

Figure 11. (A) Alignment and (B) phylogenetic relationships of archetypal HMEI-A and HMEI-B (from *H. manillensis*) and those from this study.

Saratin was an inhibitor of collagen-platelet interaction [109,110]. By competitively inhibiting the binding of von Willebrand factor (vWF) to collagen [111], the protein showed anti-platelet aggregation activity and helped prevent blood clotting [112]. Saratin was first isolated from *H. medicinails* [113], but another protein called leech antiplatelet protein (LAPP) which was found almost simultaneously from the leech *H. officinalis* [114], was suggested to be functionally homologous to saratin [30]. Recently, an additional homology was found in *H. nipponia* [112]. Two identical saratin genes were found in the *H. manillensis* genome. Similar to hirudins, the saratins had a conserved signal peptide region and six cysteine sites (Figure 12). The sequence similarity between the archetypal saratin (PDB: 2K13) and those identified in this study was 84.69%.

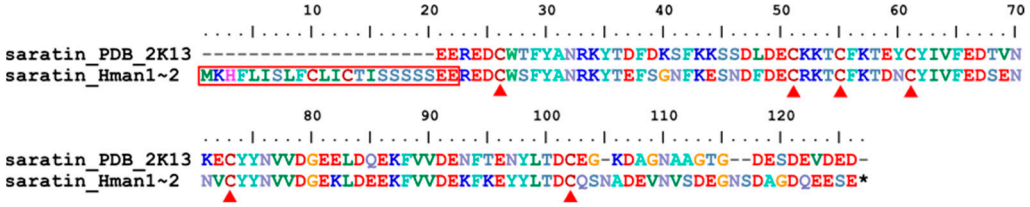

Figure 12. Alignment of archetypal saratin (from *H. officinalis*) and those from this study.

Apyrase (adenosine 5'-diphosphate diphosphohydrolase) is a nonspecific inhibitor of platelet aggregation by acting on adenosine 5'-diphosphate, arachidonic acid, platelet-activating factor, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

epinephrine. Apyrase activity is thought to be responsible for the inhibition of ADP-induced platelet aggregation, as indicated by the apparent release of apyrase from salivary glands into the host during blood feeding [115]. The bioactivities of apyrase have been studied using salivary extract from the Nile leech *Limnatis nilotica* [116], but sequence information was only available for a non-haemophagous leech (*H. robusta*, XP_009028854.1). Here we identified five apyrase proteins from the *H. manillensis* genome. One signal peptide region and two conserved cysteine sites were detected (Figure 13). The sequence similarities between the *H. robusta* apyrase (XP_009028854.1) and the six newly obtained proteins were 58.64% ~ 81.71%.

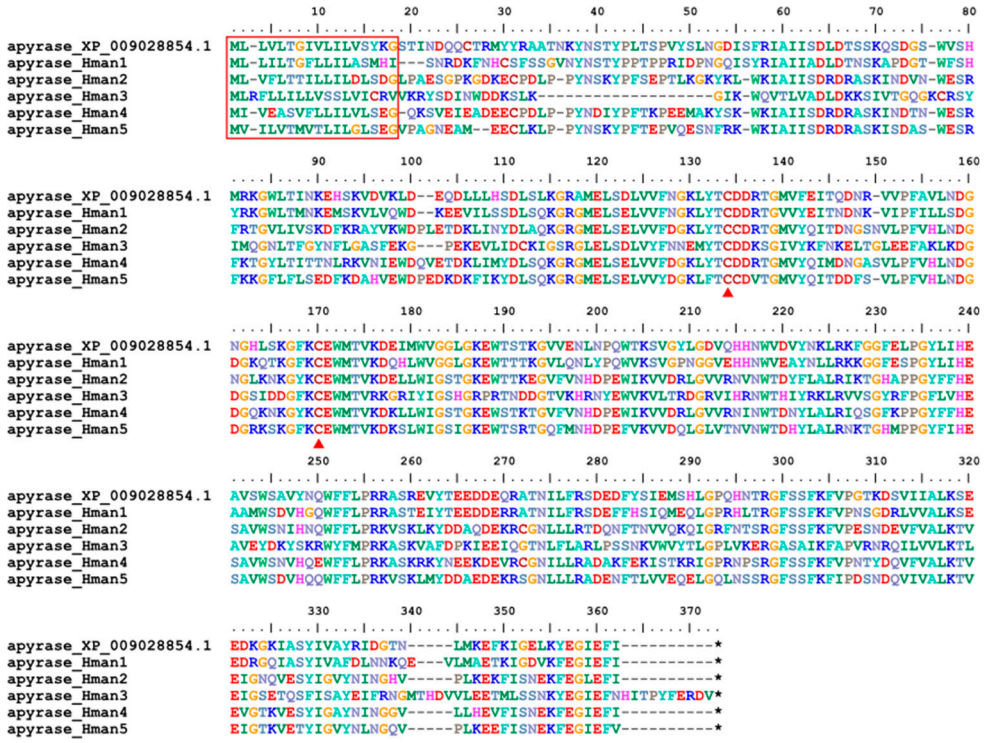

Figure 13. Alignment of archetypal apyrase (from *H. robusta*) and those from this study.

Lumbrokinase is a fibrinolytic serine protease enzyme with potent fibrinolytic activity and inhibitory effects on platelet aggregation [117]. Lumbrokinase was first extracted from the earthworm *L. rubellus* [118] and later from other earthworm species such as *Eisenia fetida* [119]. Here we report for the first-time lumbrokinase from leeches. Three forms of lumbrokinase were detected in the *H. manillensis* genome. These proteins had high sequence similarity to the archetypal lumbrokinase from *L. rubellus* (GenBank No. AAN28692.1). All leech lumbrokinases and the archetypal lumbrokinase from *L. rubellus* were relatively conserved in the signal peptide region and 14 cysteine residues (Figure 14), suggesting that these proteins also play important roles in platelet aggregation. The sequence similarities between the archetypal lumbrokinase (AAN28692.1) and those from *H. manillensis* were 67.68% ~ 69.20%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

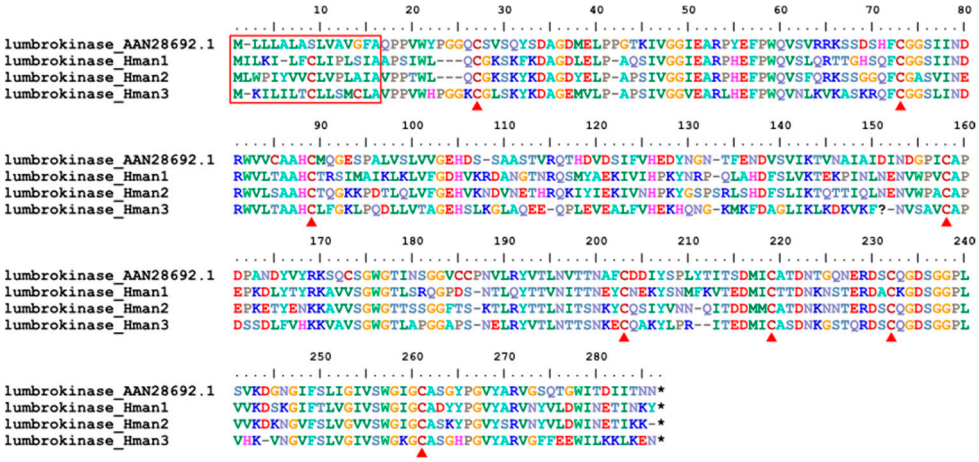

Figure 14. Alignment of archetypal lumbrokinase (from *L. rubellus*) and those from this study.

Destabilase is an endo-e(c-Glu)-Lys isopeptidase that cleaves isopeptide bonds formed by transglutaminase (Factor XIIIa) between glutamine c-carboxamide and the e-amino group of lysine. The molecule can disrupt covalent bonds formed between fibrin monomers under the influence of Factor XIIIa in blood plasma [120]. Destabilase was first detected in the salivary secretions of *H. medicinalis* [121], and three forms have been identified [122]. Similar to *H. medicinalis*, three forms were also detected in the *H. manillensis* genome. The corresponding protein sequences were relatively conserved, with the same signal region length and 14 cysteine residues (Figure 15). Surprisingly, the C-terminus of destabilase_Hman3 was extremely elongated, making its total length more than twice that of the other forms. The elongated sequences had many irregular short repeats such as "ESP" and "EST" (Figure 15). The sequence similarities between the archetypal destabilase (AAA96143.1) and those identified in this study were 70.08% ~ 73.60%.

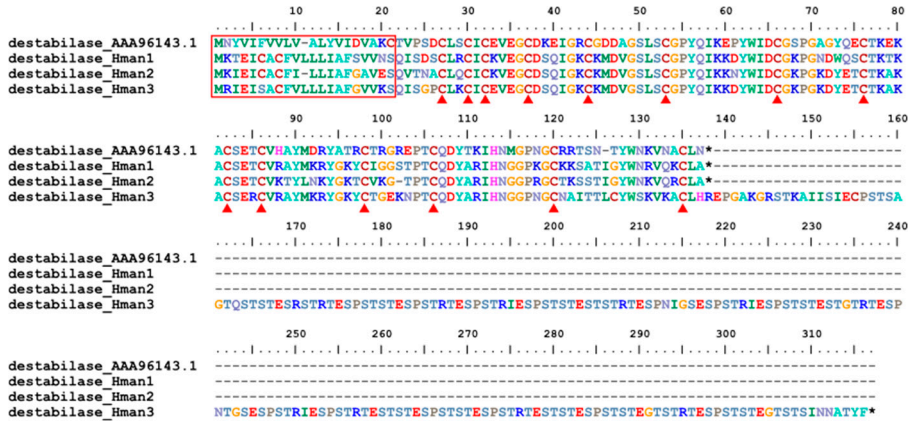

Figure 15. Alignment of archetypal destabilase (from *H. medicinalis*) and those from this study.

Gamma-glutamyl transpeptidase (GGT) is a cell surface enzyme that hydrolyzes the gamma-glutamyl bond of extracellular reduced and oxidized glutathione, initiating its cleavage into glutamate, cysteine (cystine), and glycine. It is a critical enzyme in maintaining cellular redox homeostasis and is used as a marker of liver disease and cancer [123]. A homology of mammalian GGT was isolated from the salivary gland secretion of *H. medicinalis*. This leech GGT was observed to have proteolytic activities on factor XIIIa cross-linked fibrin [124]. A GGT gene was found in the genome of *H. manillensis*. Six cysteines but no signal peptide region was detected in this protein (Figure 16). The sequence similarity between the archetypal GGT [124] and those identified in this study was 91.78%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

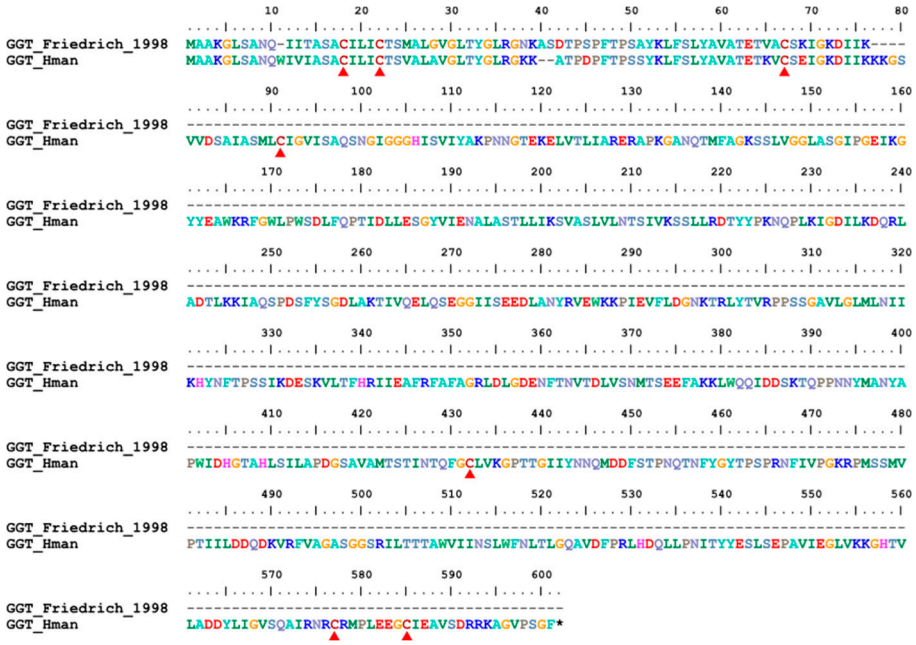

Figure 16. Alignment of archetypal GGT (from *H. medicinalis*) and those from this study.

Leech carboxypeptidase inhibitor (LCI) is a metallo-carboxypeptidase inhibitor isolated from the medicinal leech *H. medicinalis*. LCI binds tightly to pancreatic carboxypeptidases A1, A2, B and to plasma carboxypeptidases B. Assuming that leeches secrete LCI during feeding, the inhibitor could help maintain the fluid state of blood by inhibiting thrombin-activatable fibrinolysis inhibitor [125]. Here, a single LCI gene was detected in the *H. manillensis* genome. The corresponding protein had a signal peptide region and eight cysteines (Figure 17), which are critical for the structure and functional integrity of this protein [126]. After alignment, the archetypal LCI of *H. medicinalis* perfectly matched the C-terminal of the LCI of *H. manillensis*. The sequence similarity between the archetypal LCI [28] and that of *H. manillensis* was 86.89%.

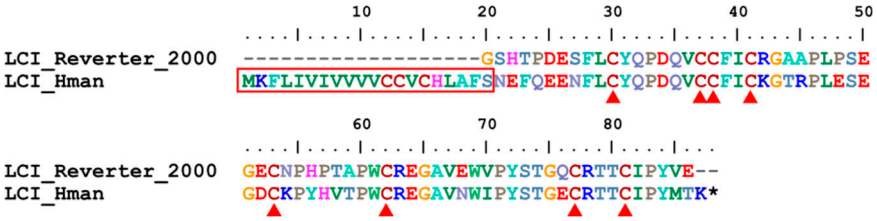

Figure 17. Alignment of archetypal LCI (from *H. medicinalis*) and that from this study.

Hyaluronidase is a spreading or diffusing substance that modifies the permeability of connective tissue by hydrolysis of endoglucoronidic bonds of hyaluronic acid [127]. Leech hyaluronidase, first isolated from *H. medicinalis*, was the most specific enzyme known for the identification of hyaluronic acid [128]. After the bite, leeches immediately release hyaluronidase to facilitate tissue penetration and spread of their bioactive molecules [129]. It has been suggested that there are three types of leech hyaluronidase [130], although sequence information is scarce. Here, for the first time, we report all three hyaluronidase genes of *H. manillensis*. One corresponding protein (hyaluronidase_Hman1), but not the other two, had a signal peptide region. Two conserved cysteines were detected in these proteins (Figure 18). The similarities between the archetypal hyaluronidase (from *H. nipponia*, AHV78514.1) and the newly obtained six proteins were 62.39% ~ 97.55%.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

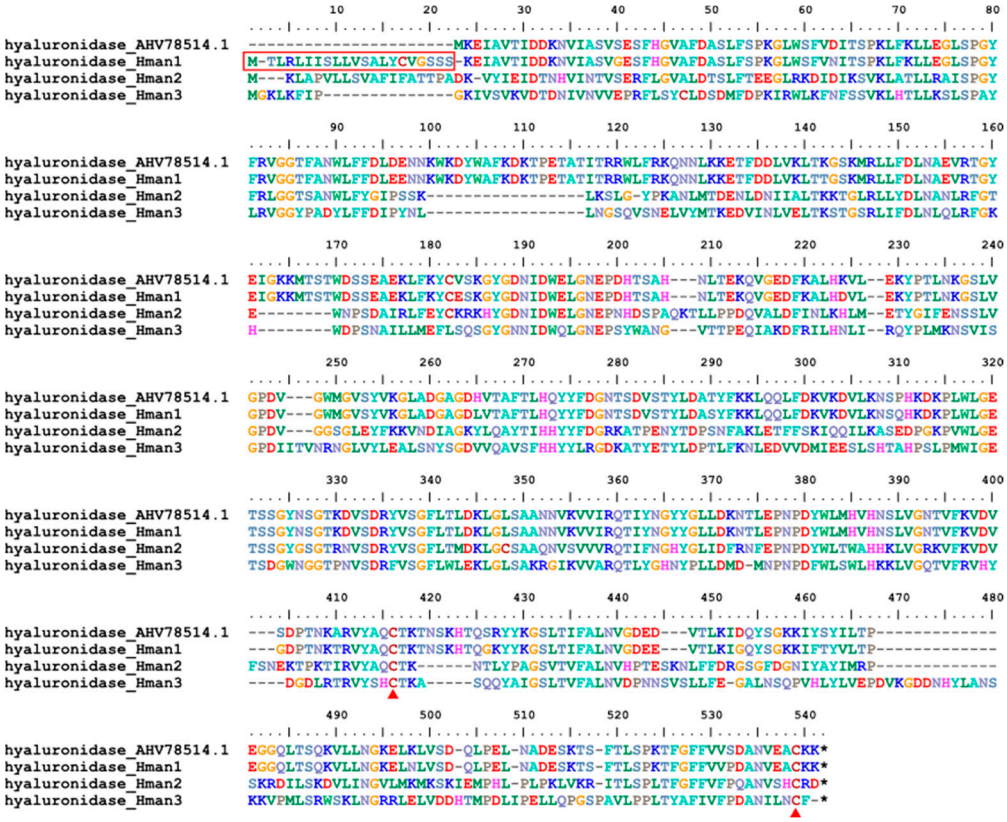

Figure 18. Alignment of archetypal hyaluronidase (from *H. nipponia*) and those from this study.

3.5. Anticoagulation of recombinant hirudins

Hirudin is the first identified and most concerned antithrombotic related gene. Here, we use *P. pastoris* as a eukaryotic expression vector to produce recombinant hirudins. The results showed that hirudin_Hman1, hirudin_Hman2, and hirudin_Hman5 exhibited obvious anticoagulant activities with mean ± SD of 960.1 ± 142.4, 508.8 ± 12.4, and 803.0 ± 101.3 U/mg, respectively. In contrast, hirudin_Hman3 and hirudin_Hman4 and the control group showed no anticoagulant activity.

We collected all available protein sequences whose anticoagulant activity had been tested and combined with the five hirudins identified in this study, and reconstructed a phylogenetic tree using IQ-TREE. The results showed that the proteins with and without anticoagulant activity clustered into distinct clades. In particular, active hirudin_Hman1 and hirudin_Hman5 were distributed in a monophyletic clade (clade A), which included seven active hirudins from *H. medicinalis* (HV1), *H. manillensis* (HM_P6, HM2, HM3, and HM4), *H. nipponia* (hirudinHN), and *W. pigra* (Wpig_V1). The active hirudin_Hman2 was distributed in another monograph (group B1), which included two active hirudins from *H. manillensis* (HLF5 and HLF8). In contrast, the inactive hirudin_Hman3 and hirudin_Hman4 were distributed in another monograph (group B2) with inactive hirudins from *H. manillensis* (HLF7a and HLF6) (Figure 19).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 November 2023 doi:10.20944/preprints202311.0026.v1

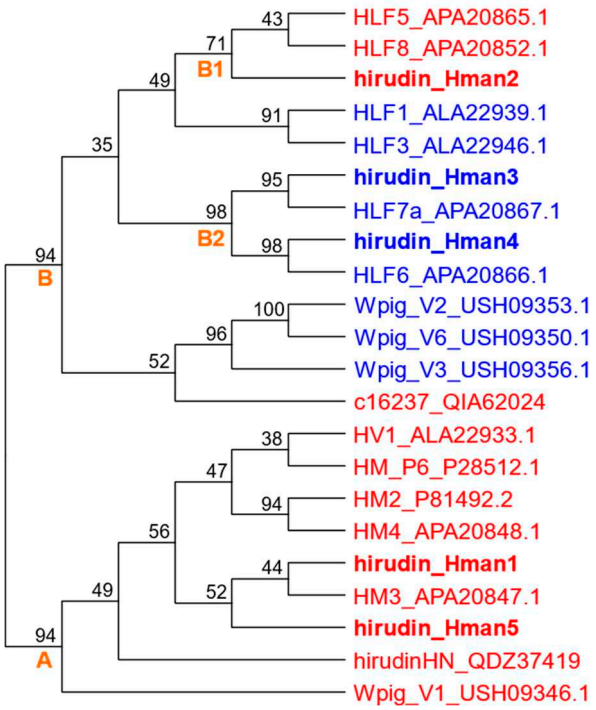

Figure 19. Phylogenetic relationship of the five hirudins from this study and all available protein sequences whose anticoagulant activity had been tested (red, active proteins; blue, inactive proteins).

4. Discussion

A high-quality genome is essential for effective identification and analysis of functional proteins. Recent high-throughput sequencing and assembly technologies make it easier to obtain a chromosome-scale genome. Here, we used state-of-the-art third-generation sequencing (PacBio HiFi) and next-generation sequencing (Illumina Hi-C, Survey and RNASeq) to obtain a nearly complete chromosome-scale genome of *H. manillensis*. A total of 13 chromosomes were identified, consistent with the recent study [35]. However, the chromosomes in our study account for 99.22% of the total length of the scaffolds, which is much higher than that in the recent study (95.92% in [35]). The BUSCO analyses showed that 95.6% of complete BUSCOs were captured. In contrast, after reanalyzing the previously published genomes [35], only 88.7% complete BUSCOs were captured for the *H. manillensis* genome. The Merqury analyses yielded a quality score of 33.5, which is also higher than that of Zheng et al. [35] (32.1). In addition, we assembled a complete circular mitochondrial genome for this species. As a result, based on the above parameters, it is almost certain that we have obtained the most integrated genome for *H. manillensis* to date.

In contrast to the high quality of the genome assemblies, the results of genetic prediction of antithrombotic genes were not very satisfactory. To test the prediction quality of different methods, we used the finally identified hirudins as query sequences to BLAST (tblastn) the CDS data sets predicted by these methods. Among the five hirudins, 2 ~ 3 hirudins were successfully predicted by GlimmerHMM, SNAP, and Braker, while none were predicted by PASA or Stringtie. Unexpectedly, no hirudin gene was found in the consensus prediction results generated by EVidenceModeler. It should be noted that the prediction of antithrombotic genes (at least for hirudins) does not necessarily seem to be better with higher completeness of genomes than in previous studies [35,42,81], indicating that formalistic prediction by program pipelines is not sufficient to identify antithrombotic genes.

Using a so-called BRAKER-plus strategy, combining BRAKER prediction and manual prediction, we finally identified 72 genes from 21 gene families involved in antithrombotic functions. All protein products of these genes had more than 50% sequence similarity to their corresponding archetypal proteins, indicating that our identification methods are reliable. Except for lefaxin, eglin, apyrase, and hyaluronidase, most of the antithrombotic proteins were cysteine-rich, forming three or more disulfide bonds. By increasing protein stability, disulfide bonds were thought to make protein

 doi:10.20944/preprints202311.0026.v1

structures relatively independent of their amino acid sequences, thus acting as buffers against deleterious mutations and allowing accelerated sequence evolution [131]. Interestingly, three of the cysteine-rich antithrombotic protein families (LDTI, antistasin, and granulin) have internal tandem repeats containing 6, 10, and 12 conserved cysteines, respectively. The pattern of internal tandem oligopeptide repeats is not uncommon in natural proteins; however, the functional and evolutionary significance of such a pattern has not been well studied [132].

Leeches have been used as a medical and pharmaceutical resource for many centuries, and massive bioactive proteins have been reported, and many species (*Hirudo* spp., *Hirudinaria* spp., *Haemadipsa* spp., *Haementeria* spp., *Theromyzon* spp. and etc.) have been involved in related studies [133]. The high variation of antithrombotic proteins and their coding genes cause confusion in the research and application of these substances. Taking hirudin as an example, in addition to the archetypal hirudin identified from *H. medicinalis*, several alternative names have been used: hirudin-like factors that have no anticoagulant activity [78], haemadin from *Haemadipsa* spp. [80], bufrudin from *Hirudinaria* spp. [134]. In addition, most of the previous studies involve only a single gene of a gene family. In fact, many proteins are encoded by multigene families. In the precent study, the high-quality genomes and careful personalized analysis provide an opportunity to systematically investigate the antithrombotic proteins as well as genes of *H. manillensis*.

With respect to the three antithrombotic drugs (anticoagulation, antiplatelet aggregation, and fibrinolysis), we broadly classified the antithrombotic proteins of leeches in this study into three categories: inhibitors of coagulation, inhibitors of platelet aggregation, and enhancers of fibrinolysis. Of the 21 antithrombotic protein families identified, 14 (67%) were involved in coagulation inhibition, including two thrombin inhibitors (hirudin and granulin), three Factor Xa inhibitors (antistasin, therostasin and lefaxin), and nine serine protease inhibitors (hirustasin, guamerin, piguamerin, bdellastasin, poecistasin, eglin, bdellin, LDTI, and HMEI). Three gene families (saratin, apyrase, and lumbrokinase) were involved in the inhibition of platelet aggregation. Another three (destabilase, GGT, and LCI) were involved in enhancing fibrinolysis. The remaining family, hyaluronidase, does not belong to these three categories, but since it facilitates tissue penetration and diffusion of the other 20 protein families, its indirect antithrombotic function is not negligible.

Hirudin is the most potent thrombin-specific inhibitor identified to date and is a representative pharmacologically active substance in leeches [25,135]. In the Pharmacopoeia of the People's Republic of China (PPRC), antithrombin activity is the key index for determining the quality of leeches [73]. The anticoagulation analysis of *H. manillensis* hirudins in this study showed that three (hirudin_Hman1, hirudin_Hman2, and hirudin_Hman5) out of five hirudins had anticoagulant activity. We also collected the sequence and functional information of hirudins from previous reports, and then tested the phylogenetic relationships of the reported hirudins and those from this study. The results showed that the proteins with and without anticoagulant activity clustered into separate clades. For example, the active hirudin_Hman1, hirudin_Hman2, and hirudin_Hman5 had closer relationships with proteins that were confirmed to have anticoagulant activity. In contrast, the inactive hirudin_Hman3 and hirudin_Hman4 were more phylogenetically related to proteins confirmed to have no anticoagulant activity. These results repeatedly confirmed that at least three functionally active hirudins were simultaneously distributed in a single *H. manillensis* genome.

In conclusion, in this study we provide an almost complete, high-quality genome of *H. manillensis*. Combined with automatic and manual prediction, we identified 72 antithrombotic genes involving 21 gene families. The functions of the corresponding proteins include anticoagulation, anti-platelet aggregation, fibrinolysis, and drug diffusion. We have also provided the complete CDS and protein sequences and their variation information for all 72 antithrombotic genes/proteins. This is the most comprehensive collection of genomes and leech antithrombotic biomacromolecules to date. Our results will greatly facilitate the research and application of leech derivatives for medical and pharmaceutical purposes of thrombosis.

Supplementary Materials: The following supporting information can be downloaded at the Figshare online repository: https://doi.org/10.6084/m9.figshare.24187377.v1. These are the genome assembly

(1Hman.genome.fa), GFF file of BRAKER-plus gene prediction approach (2Hman.genome.gff), all predicted CDS (3Hman.cds.fa), and all CDS of 72 antithrombotic genes (4Hman.antithrombotic.cds.fa) of *H. manillensis*.

Author Contributions: Conceptualization, Z.L., F.Z., and G.L.; methodology, B.H., Y.L., and G.L.; software, G.L.; validation, formal analysis, investigation, resources, data curation, visualization, B.H., Y.L., Q.H., R.M., J.Q., and G.L.; writing—original draft preparation, writing—review and editing, Z.L., F.Z., and G.L.; supervision, project administration and funding acquisition, Z.H., F.Z., and G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (No. 82260742 and 32260132), the Jiangxi "Double Thousand Plan" (No. jxsq2020101050) , the Yunnan Provincial Ten Thousand People Plan (YNWRQNBJ-2018-161), and the Frontier Research Team of Kunming University 2023.

Data Availability Statement: The sequence data (clean reads) of PacBio, Hi-C, Survey, RNASeq were deposited in GenBank under Project PRJNA1019887.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. DeLoughery, T.G. Hemostasis and thrombosis. Springer Nature Switzerland AG, Switzerland, 2019.
2. WHO. The top 10 causes of death. 2020.
3. Mackman, N.; Bergmeier, W.; Stouffer, G.A.; Weitz, J.I. Therapeutic strategies for thrombosis: new targets and approaches. *Nat Rev Drug Discov* **2020**, *19*, 333-352.
4. Elantably, D.; Mourad, A.; Elantably, A.; Effat, M. Warfarin induced leukocytoclastic vasculitis: an extraordinary side effect. *J Thromb Thrombolysis* **2020**, *49*, 149-152.
5. Cheng, Y.J.; Wang, Y.N.; Song, Q.H.; Qiu, K.; Liu, M. Use of anticoagulant therapy and cerebral microbleeds: a systematic review and meta-analysis. *J Neurol* **2021**, *268*, 1666-1679.
6. Al-Husein, B.A.; Al-Azzam, S.I.; Alzoubi, K.H.; Khabour, O.F.; Nusair, M.B.; Alzayadeen, S. Investigating the effect of demographics, clinical characteristics, and polymorphism of MDR-1, CYP1A2, CYP3A4, and CYP3A5 on clopidogrel resistance. *J Cardiovasc Pharmacol* **2018**, *72*, 296-302.
7. Giahchi, F.; Mohammadi, M. Reteplase versus streptokinase in management of ST-segment elevation myocardial infarction; a letter to the editor. *Adv J Emerg Med* **2019**, *3*, e34.
8. Sawyer, R.T. Leech biology and behaviour. Oxford University Press, Oxford, UK, 1986.
9. Kvist, S.; Manzano-Marín, A.; de Carle, D.; Trontelj, P.; Siddall, M.E. Draft genome of the European medicinal leech *Hirudo medicinalis* (Annelida, Clitellata, Hirudiniformes) with emphasis on anticoagulants. *Sci Rep* **2020**, *10*, 9885.
10. Sket, B.; Trontelj, P. Global diversity of leeches (Hirudinea) in freshwater. *Hydrobiologia* **2008**, *595*, 129-137.
11. Sig, A.K.; Guney, M.; Guclu, A.U.; Ozmen, E. Medicinal leech therapy—an overall perspective. *Integr Med Res* **2017**, *6*, 337-343.
12. Ma, C.J.; Li, X.; Chen, H. Research progress in the use of leeches for medical purposes. *Tradit Med Res* **2021**, *6*, 56-69.
13. Shakouri, A.; Adljouy, N.; Balkani, S.; Mohamadi, M.; Hamishehkar, H.; Abdolalizadeh, J.; Shakouri, S.K. Effectiveness of topical gel of medical leech (*Hirudo medicinalis*) saliva extract on patients with knee osteoarthritis: a randomized clinical trial. *Complement Ther Clin Pract* **2018**, *31*, 352-359.
14. Hohmann, C.D.; Stange, R.; Steckhan, N.; Robens, S.; Ostermann, T.; Paetow, A.; Michalsen, A. The effectiveness of leech therapy in chronic low back pain. *Dtsch Arztebl Int* **2018**, *115*, 785-792.
15. Hamidizadeh, N.; Azizi, A.; Zarshenas, M.M.; Ranjbar, S. Leech therapy in treatment of cutaneous leishmaniasis: a case report. *J Integr Med* **2017**, *15*, 407-410.
16. Michalsen, A.; Roth, M.; Dobos, G. Medicinal leech therapy. Georg Thieme Verlag, Stuttgart, Germany, 2007.
17. Elyassi, A.R.; Terres, J.; Rowshan, H.H. Medicinal leech therapy on head and neck patients: a review of literature and proposed protocol. *Oral Surg Oral Med Oral Pathol Oral Radiol* **2013**, *116*, e167-e172.
18. Dong, H.; Ren, J.X.; Wang, J.J.; Ding, L.S.; Zhao, J.J.; Liu, S.Y.; Gao, H.M. Chinese medicinal leech: ethnopharmacology, phytochemistry, and pharmacological activities. *Evid Based Complement Alternat Med* **2016**, *2016*, 7895935.
19. Song, J.X.; Lyu, Y.; Wang, M.M.; Zhang, J.; Gao, L.; Tong, X.L. Treatment of human urinary kallidinogenase combined with maixuekang capsule promotes good functional outcome in ischemic stroke. *Front Physiol* **2018**, *9*, 84.
20. Zaidi, S.M.A.; Jameel, S.S.; Zaman, F.; Jilani, S.; Sultana, A.; Khan, S.A. A systematic overview of the medicinal importance of sanguivorous leeches. *Altern Med Rev* **2011**, *16*, 59-65.
21. Fritsma, G.A. Monitoring the direct thrombin inhibitors. *Clin Lab Sci* **2013**, *26*, 54-57.

22. Müller, C.; Lukas, P.; Böhmert, M.; Hildebrandt, J.P. Hirudin or hirudin-like factor—that is the question: insights from the analyses of natural and synthetic HLF variants. *FEBS Lett* **2020**, *594*, 841-850.
23. Mousa, R.; Hidmi, T.; Pomyalov, S.; Lansky, S.; Khouri, L.; Shalev, D.E.; Shoham, G.; Metanis, N. Diselenide crosslinks for enhanced and simplified oxidative protein folding. *Commun Chem* **2021**, *4*, 30.
24. Montinari, M.R.; Minelli, S. From ancient leech to direct thrombin inhibitors and beyond: New from old. *Biomed Pharmacother* **2022**, *149*, 112878.
25. Chen, J.R.; Xie, X.F.; Zhang, H.Q.; Li, G.M.; Yin, Y.P.; Cao, X.Y.; Gao, Y.Q.; Li, Y.N.; Zhang, Y.; Peng, F.; et al. Pharmacological activities and mechanisms of hirudin and its derivatives—a review. *Front Pharmacol* **2021**, *12*, 660757.
26. Cheng, R.M.; Tang, X.P.; Long, A.L.; Mwangi, J.; Lai, R.; Sun, R.P.; Long, C.B.; Zhang, Z.Q. Purification and characterization of a novel anti-coagulant from the leech *Hirudinaria manillensis*. *Zool Res* **2019**, *40*, 205-210.
27. Zavalova, L.; Lukyanov, S.; Baskova, I.; Snezhkov, E.; Akopov, S.; Berezhnoy, S.; Bogdanova, E.; Barsova, E.; Sverdlov, E.D. Genes from the medicinal leech (*Hirudo medicinalis*) coding for unusual enzymes that specifically cleave endo-epsilon (gamma-Glu)-Lys isopeptide bonds and help to dissolve blood clots. *Mol Gen Genet* **1996**, *253*, 20-25.
28. Reverter, D.; Fernández-Catalán, C.; Baumgartner, R.; Pfänder, R.; Huber, R.; Bode, W.; Vendrell, J.; Holak, T.A.; Avilés, F.X. Structure of a novel leech carboxypeptidase inhibitor determined free in solution and in complex with human carboxypeptidase A2. *Nat Struct Biol* **2000**, *7*, 322-328.
29. Jin, P.; Kang, Z.; Zhang, N.; Du, G.C.; Chen, J. High-yield novel leech hyaluronidase to expedite the preparation of specific hyaluronan oligomers. *Sci Rep* **2014**, *4*, 4471.
30. Gronwald, W.; Bomke, J.; Maurer, T.; Domogalla, B.; Huber, F.; Schumann, F.; Kremer, W.; Fink, F.; Rysiok, T.; Frech, M.; et al. Structure of the leech protein saratin and characterization of its binding to collagen. *J Mol Biol* **2008**, *381*, 913-927.
31. Kvist, S.; Min, G.S.; Siddall, M.E. Diversity and selective pressures of anticoagulants in three medicinal leeches (Hirudinida: Hirudinidae, Macrobdellidae). *Ecol Evol* **2013**, *3*, 918-933.
32. Iwama, R.E.; Tessler, M.; Kvist, S. Leech anticoagulants are ancestral and likely to be multifunctional. *Zool J Linn Soc* **2022**, *196*, 137-148.
33. Babenko, V.V.; Podgorny, O.V.; Manuvera, V.A.; Kasianov, A.S.; Manolov, A.I.; Grafskaia, E.N.; Shirokov, D.A.; Kurdyumov, A.S.; Vinogradov, D.V.; Nikitina, A.S.; et al. Draft genome sequences of *Hirudo medicinalis* and salivary transcriptome of three closely related medicinal leeches. *BMC Genomics* **2020**, *21*, 331.
34. Zheng, F.S.; Zhang, M.; Yang, X.W.; Wu, F.L.; Wang, G.; Feng, X.X.; Ombati, R.; Zuo, R.L.; Yang, C.J.; Liu, J.; et al. Prostaglandin E1 is an efficient molecular tool for forest leech blood sucking. *Front Vet Sci* **2021**, *7*, 615915.
35. Zheng, J.H.; Wang, X.B.; Feng, T.; Rehman, S.U.; Yan, X.Y.; Shan, H.Q.; Ma, X.C.; Zhou, W.G.; Xu, W.H.; Lu, L.Y.; et al. Molecular mechanisms underlying hematophagia revealed by comparative analyses of leech genomes. *Gigascience* **2023**, *12*, 1-11.
36. Müller, C.; Haase, M.; Lemke, S.; Hildebrandt, J.P. Hirudins and hirudin-like factors in Hirudinidae: implications for function and phylogenetic relationships. *Parasitol Res* **2017**, *116*, 313-325.
37. Tubtimon, J.; Jeratthitikul, E.; Sutcharit, C.; Kongim, B.; Panha, S. Systematics of the freshwater leech genus *Hirudinaria* Whitman, 1886 (Arhynchobdellida, Hirudinidae) from northeastern Thailand. *Zookeys* **2014**, *452*, 15-33.
38. Jeratthitikul, E.; Jiranuntskul, P.; Nakano, T.; Sutcharit, C.; Panha, S. A new species of buffalo leech in the genus *Hirudinaria* Whitman, 1886 (Arhynchobdellida, Hirudinidae) from Thailand. *Zookeys* **2020**, *933*, 1-14.
39. Electricwala, A.; Hartwell, R.; Scawen, M.D.; Atkinson, T. The complete amino acid sequence of a hirudin variant from the leech *Hirudinaria manillensis*. *J Protein Chem* **1993**, *12*, 365-370.
40. Liu, F.; Guo, Q.S.; Shi, H.Z.; Cheng, B.X.; Lu, Y.X.; Gou, L.; Wang, J.; Shen, W.B.; Yan, S.M.; Wu, M.J. Genetic variation in *Whitmania pigra*, *Hirudo nipponica* and *Poecilobdella manillensis*, three endemic and endangered species in China using SSR and TRAP markers. *Gene* **2016**, *579*, 172-182.
41. Zhang, B.; Wang, B.; Gong, Y.; Yu, X.; Lv, J.Y. Anticoagulant active substances extraction and anti-thrombin activity analysis of several species of leeches. *Acta Sci Nat Univ Sun* **2012**, *51*, 92-96.
42. Guan, D.L.; Yang, J.; Liu, Y.K.; Li, Y.; Mi, D.; Ma, L.B.; Wang, Z.Z.; Xu, S.Q.; Qiu, Q. Draft genome of the Asian buffalo leech *Hirudinaria manillensis*. *Front Genet* **2020**, *10*, 1321.
43. Hu, J.; Fan, J.; Sun, Z.; Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **2020**, *36*, 2253-2255.
44. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.
45. Zhang, H.W.; Song, L.; Wang, X.T.; Cheng, H.Y.; Wang, C.F.; Meyer, C.A.; Liu, T.; Tang, M.; Aluru, S.; Yue, F.; et al. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nat Commun* **2021**, *12*, 6566.

46. Durand, N.C.; Shamim, M.S.; Machol, I.; Rao, S.S.P.; Huntley, M.H.; Lander, E.S.; Aiden, E.L. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **2016a**, *3*, 95-98.
47. Durand, N.C.; Robinson, J.T.; Shamim, M.S.; Machol, I.; Mesirov, J.P.; Lander, E.S.; Aiden, E.L. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **2016b**, *3*, 99-101.
48. Jin, J.J.; Yu, W.B.; Yang, J.B.; Song, Y.; dePamphilis, C.W.; Yi, T.S.; Li, D.Z. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. *Genome Biol* **2020**, *21*, 241.
49. Seppey, M.; Manni, M.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness. *Methods Mol Biol* **2019**, *1962*, 227-245.
50. Rhie, A.; Walenz, B.P.; Koren, S.; Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **2020**, *21*, 245.
51. Flynn, J.M.; Hubley, R.; Goubert, C.; Rosen, J.; Clark, A.G.; Feschotte, C.; Smit, A.F. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **2020**, *117*, 9451-9457.
52. Bao, W.D.; Kojima, K.K.; Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **2015**, *6*, 11.
53. Majoros, W.H.; Pertea, M.; Salzberg, S.L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **2004**, *20*, 2878-2879.
54. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **2004**, *5*, 59.
55. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.D.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **2011**, *29*, 644-652.
56. Haas, B.J.; Delcher, A.L.; Mount, S.M.; Wortman, J.R.; Smith, R.K.J.; Hannick, L.I.; Maiti, R.; Ronning, C.M.; Rusch, D.B.; Town, C.D.; et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **2003**, *31*, 5654-5666.
57. Kim, D.; Langmead, B.; Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **2015**, *12*, 357-360.
58. Pertea, M.; Pertea, G.M.; Antonescu, C.M.; Chang, T.C.; Mendell, J.T.; Salzberg, S.L. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **2015**, *33*, 290-295.
59. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15-21.
60. Hoff, K.J.; Lange, S.; Lomsadze, A.; Borodovsky, M.; Stanke, M. BRAKER1: Unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **2016**, *32*, 767-769.
61. Haas, B.J.; Salzberg, S.L.; Zhu, W.; Pertea, M.; Allen, J.E.; Orvis, J.; White, O.; Buell, C.R.; Wortman, J.R. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol* **2008**, *9*, R7.
62. Pertea, G.; Pertea, M. GFF utilities: GffRead and GffCompare. *F1000Res* **2020**, *9*, ISCB Comm J-304.
63. Sharma, K.; Singh, A.K.; Maddipatla, D.K.; Deshwal, G.K.; Rao, P.S.; Sharma, H. Eggnog: process optimization and characterization of a dairy-based beverage. *J Dairy Res* **2023**, *90*, 205-212.
64. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* **2021**, *49*, D412-D419.
65. Tang, S.; Lomsadze, A.; Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res* **2015**, *43*, e78.
66. Slater, G.S.C.; Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **2005**, *6*, 31.
67. Dainat, J. AGAT: Another gff analysis toolkit to handle annotations in any GTF/GFF format. (v0.7.0). *Zenodo* **2021**, https://www.doi.org/10.5281/zenodo.3552717.
68. Shen, W.; Le, S.; Li, Y.; Hu, F.Q. SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PLoS One* **2016**, *11*, e0163962.
69. Kumar, S.; Stecher, G.; Li, M.; Knyaz, C.; Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* **2018**, *35*, 1547-1549.
70. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: the European molecular biology open software suite. *Trends Genet* **2000**, *16*, 276-277.
71. Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **2015**, *32*, 268-274.
72. Hu, Z.L.; Zhang, N.; Gu, F.; Li, Y.; Deng, X.J.; Chen, G.P. Expression, purification and characterization of recombinant targeting bifunctional hirudin in *Pichia pastoris*. *Afr J Biotechnol* **2009**, *8*, 5582-5588.
73. Commission, C.P. Pharmacopoeia of the people's republic of China. Medicine Science and Technology Press, Beijing, China, 2020.
74. Markwardt, F. Untersuchungen über hirudin. *Naturwissenschaften* **1955**, *42*, 537-538.

75. Vitali, J.; Martin, P.D.; Malkowski, M.G.; Robertson, W.D.; Lazar, J.B.; Winant, R.C.; Johnson, P.H.; Edwards, B.F. The structure of a complex of bovine alpha-thrombin and recombinant hirudin at 2.8-A resolution. *J Biol Chem* **1992**, *267*, 17670-17678.
76. Dodt, J.; Müller, H.P.; Seemüller, U.; Chang, J.Y. The complete amino acid sequence of hirudin, a thrombin specific inhibitor: application of colour carboxymethylation. *FEBS Lett* **1984**, *165*, 180-184.
77. Harvey, R.P.; Degryse, E.; Stefani, L.; Schamber, F.; Cazenave, J.P.; Courtney, M.; Tolstoshev, P.; Lecocq, J.P. Cloning and expression of a cDNA coding for the anticoagulant hirudin from the bloodsucking leech, *Hirudo medicinalis*. *Proc Natl Acad Sci U S A* **1986**, *83*, 1084-1088.
78. Müller, C.; Mescke, K.; Liebig, S.; Mahfoud, H.; Lemke, S.; Hildebrandt, J.P. More than just one: multiplicity of hirudins and hirudin-like factors in the medicinal leech, *Hirudo medicinalis*. *Mol Genet Genomics* **2016**, *291*, 227-240.
79. Scacheri, E.; Nitti, G.; Valsasina, B.; Orsini, G.; Visco, C.; Ferrera, M.; Sawyer, R.T.; Sarmientos, P. Novel hirudin variants from the leech *Hirudinaria manillensis*. Amino acid sequence, cDNA cloning and genomic organization. *Eur J Biochem* **1993**, *214*, 295-304.
80. Strube, K.H.; Kröger, B.; Bialojan, S.; Otte, M.; Dodt, J. Isolation, sequence analysis, and cloning of haemadin. An anticoagulant peptide from the Indian leech. *J Biol Chem* **1993**, *268*, 8590-8595.
81. Tong, L.; Dai, S.X.; Kong, D.J.; Yang, P.P.; Tong, X.; Tong, X.R.; Bi, X.X.; Su, Y.; Zhao, Y.Q.; Liu, Z.C. The genome of medicinal leech (*Whitmania pigra*) and comparative genomic study for exploration of bioactive ingredients. *BMC Genomics* **2022**, *23*, 76.
82. Tanaka, Y.; Suzuki, G.; Matsuwaki, T.; Hosokawa, M.; Serrano, G.; Beach, T.G.; Yamanouchi, K.; Hasegawa, M.; Nishihara, M. Progranulin regulates lysosomal function and biogenesis through acidification of lysosomes. *Hum Mol Genet* **2017**, *26*, 969-988.
83. Hong, S.J.; Kang, K.W. Purification of granulin-like polypeptide from the blood-sucking leech, *Hirudo nipponia*. *Protein Expr Purif* **1999**, *16*, 340-346.
84. Mittl, P.R.E.; Di Marco, S.; Fendrich, G.; Pohlig, G.; Heim, J.; Sommerhoff, C.; Fritz, H.; Priestle, J.P.; Grütter, M.G. A new structural class of serine protease inhibitors revealed by the structure of the hirustasin-kallikrein complex. *Structure* **1997**, *5*, 253-264.
85. Dunwiddie, C.; Thornberry, N.A.; Bull, H.G.; Sardana, M.; Friedman, P.A.; Jacobs, J.W.; Simpson, E. Antistasin, a leech-derived inhibitor of factor Xa. Kinetic analysis of enzyme inhibition and identification of the reactive site. *J Biol Chem* **1989**, *264*, 16694-16699.
86. Nutt, E.; Gasic, T.; Rodkey, J.; Gasic, G.J.; Jacobs, J.W.; Friedman, P.A.; Simpson, E. The amino acid sequence of antistasin. A potent inhibitor of factor Xa reveals a repeated internal structure. *J Biol Chem* **1988**, *263*, 10162-10167.
87. Blankenship, D.T.; Brankamp, R.G.; Manley, G.D.; Cardin, A.D. Amino acid sequence of ghilanten: Anticoagulant-antimetastatic principle of the south American leech, *Haementeria ghilianii*. *Biochem Bioph Res Co* **1990**, *166*, 1384-1389.
88. Han, J.H.; Law, S.W.; Keller, P.M.; Kniskern, P.J.; Silberklang, M.; Tung, J.S.; Gasic, T.B.; Gasic, G.J.; Friedman, P.A.; Ellis, R.W. Cloning and expression of cDNA encoding antistasin, a leech-derived protein having anti-coagulant and anti-metastatic properties. *Gene* **1989**, *75*, 47-57.
89. Faria, F.; Kelen, E.M.; Sampaio, C.A.; Bon, C.; Duval, N.; Chudzinski-Tavassi, A.M. A new factor Xa inhibitor (lefaxin) from the *Haementeria* depressa leech. *Thromb Haemost* **1999**, *82*, 1469-1473.
90. Chopin, V.; Salzet, M.; Baert, J.l.; Vandenbulcke, F.; Sautiére, P.E.; Kerckaert, J.P.; Malecha, J. Therostasin, a novel clotting factor Xa inhibitor from the rhynchobdellid leech, *Theromyzon tessulatum*. *J Biol Chem* **2000**, *275*, 32701-32707.
91. Söllner, C.; Mentele, R.; Eckerskorn, C.; Fritz, H.; Sommerhoff, C.P. Isolation and characterisation of hirustasin, an antistasin-type serine-proteinase inhibitor from the medical leech *Hirudo medicinalis*. *Eur J Biochem* **1994**, *219*, 937-943.
92. Jung, H.I.; Kim, S.I.; Ha, K.S.; Joe, C.O.; Kang, K.W. Isolation and characterization of guamerin, a new human leukocyte elastase inhibitor from *Hirudo nipponia*. *J Biol Chem* **1995**, *270*, 13879-13884.
93. Kim, D.R.; Kang, K.W. Amino acid sequence of piguamerin, an antistasin-type protease inhibitor from the blood sucking leech *Hirudo nipponia*. *Eur J Biochem* **1998**, *254*, 692-697.
94. Kim, D.R.; Hong, S.J.; Ha, K.S.; Joe, C.O.; Kang, K.W. A cysteine-rich serine protease inhibitor (guamerin II) from the non-blood sucking leech *Whitmania edentula*: biochemical characterization and amino acid sequence analysis. *J Enzyme Inhib* **1996**, *10*, 81-91.
95. Moser, M.; Auerswald, E.; Mentele, R.; Eckerskorn, C.; Fritz, H.; Fink, E. Bdellastasin, a serine protease inhibitor of the antistasin family from the medical leech (*Hirudo medicinalis*)—Primary structure, expression in yeast, and characterisation of native and recombinant inhibitor. *Eur J Biochem* **1998**, *253*, 212-220.
96. Tang, X.P.; Chen, M.R.; Duan, Z.L.; Mwangi, J.; Li, P.P.; Lai, R. Isolation and characterization of poecistasin, an anti-thrombotic antistasin-type serine protease inhibitor from leech *Poecilobdella manillensis*. *Toxins (Basel)* **2018**, *10*, 429.

97. Seemüller, U.; Eulitz, M.; Fritz, H.; Strobl, A. Structure of the elastase-cathepsin G inhibitor of the leech *Hirudo medicinalis*. *Hoppe Seylers Z Physiol Chem* **1980**, *361*, 1841-1846.
98. Renesto, P.; Ferrer-Lopez, P.; Chignard, M. Eglin C and heparin inhibition of platelet activation induced by cathepsin G or human neutrophils. *Ann N Y Acad Sci* **1991**, *624*, 321-324.
99. Chang, J.Y.; Knecht, R.; Maschler, R.; Seemüller, U. Elastase-cathepsin G inhibitors eglin b and eglin c differ by a single Tyr----His substitution. A micro-method for the identification of amino-acid substitution. *Biol Chem Hoppe Seyler* **1985**, *366*, 281-286.
100. Fritz, H.; Oppitz, K.H.; Gebhardt, M.; Oppitz, I.; Werle, E.; Marx, R. On the presence of a trypsin-plasmin inhibitor in hirudin. *Hoppe Seylers Z Physiol Chem* **1969**, *350*, 91-92.
101. Fink, E.; Rehm, H.; Gippner, C.; Bode, W.; Eulitz, M.; Machleidt, W.; Fritz, H. The primary structure of bdellin B-3 from the leech *Hirudo medicinalis*. Bdellin B-3 is a compact proteinase inhibitor of a "non-classical" Kazal type. It is present in the leech in a high molecular mass form. *Biol Chem Hoppe Seyler* **1986**, *367*, 1235-1242.
102. Laskowski, M.J.; Kato, I. Protein inhibitors of proteinases. *Annu Rev Biochem* **1980**, *49*, 593-626.
103. Campos, I.T.N.; Silva, M.M.; Azzolini, S.S.; Souza, A.F.; Sampaio, C.A.M.; Fritz, H.; Tanaka, A.S. Evaluation of phage display system and leech-derived tryptase inhibitor as a tool for understanding the serine proteinase specificities. *Arch Biochem Biophys* **2004**, *425*, 87-94.
104. Sommerhoff, C.P.; Söllner, C.; Mentele, R.; Piechottka, G.P.; Auerswald, E.A.; Fritz, H. A Kazal-type inhibitor of human mast cell tryptase: isolation from the medical leech *Hirudo medicinalis*, characterization, and sequence analysis. *Biol Chem Hoppe Seyler* **1994**, *375*, 685-694.
105. Pohlig, G.; Fendrich, G.; Knecht, R.; Eder, B.; Piechottka, G.; Sommerhoff, C.P.; Heim, J. Purification, characterization and biological evaluation of recombinant leech-derived tryptase inhibitor (rLDTI) expressed at high level in the yeast *Saccharomyces cerevisiae*. *Eur J Biochem* **1996**, *241*, 619-626.
106. Tanaka, A.S.; Silva, M.M.; Torquato, R.J.; Noguti, M.A.; Sampaio, C.A.; Fritz, H.; Auerswald, E.A. Functional phage display of leech-derived tryptase inhibitor (LDTI): construction of a library and selection of thrombin inhibitors. *FEBS Lett* **1999**, *458*, 11-16.
107. Xu, K.H.; Zhou, M.; Wu, F.L.; Tang, X.P.; Lu, Q.M.; Lai, R.; Long, C.B. Identification and characterization of a novel elastase inhibitor from *Hirudinaria manillensis*. *Chin J Nat Med* **2021**, *19*, 540-544.
108. Zhou, Y.L.; Xu, Z.D.; Liu, Z.Q. Impact of neutrophil extracellular traps on thrombosis formation: new findings and future perspective. *Front Cell Infect Microbiol* **2022**, *12*, 910908.
109. Smith, T.P.; Alshafie, T.A.; Cruz, C.P.; Fan, C.Y.; Brown, A.T.; Wang, Y.; Eidt, J.F.; Moursi, M.M. Saratin, an inhibitor of collagen-platelet interaction, decreases venous anastomotic intimal hyperplasia in a canine dialysis access model. *Vasc Endovascular Surg* **2003**, *37*, 259-269.
110. White, T.C.; Bernyl, M.A.; Robinson, D.K.; Yin, H.; DeGrado, W.F.; Hanson, S.R.; McCarthy, O.J.T. The leech product saratin is a potent inhibitor of platelet integrin $\alpha 2\beta 1$ and von Willebrand factor binding to collagen. *FEBS J* **2007**, *274*, 1481-1491.
111. Cruz, C.P.; Eidt, J.; Drouilhet, J.; Brown, A.T.; Wang, Y.; Barnes, C.S.; Moursi, M.M. Saratin, an inhibitor of von Willebrand factor-dependent platelet adhesion, decreases platelet aggregation and intimal hyperplasia in a rat carotid endarterectomy model. *J Vasc Surg* **2001**, *34*, 724-729.
112. Cheng, B.X.; Kuang, S.T.; Shao, G.Y.; Tian, Q.Q.; Gao, T.Y.; Che, X.F.; Ao, H.W.; Zhang, K.; Liu, F. Molecular cloning and functional analysis of HnSaratin from *Hirudo nipponia*. *Gene* **2023**, *869*, 147401.
113. Barnes, C.S.; Krafft, B.; Frech, M.; Hofmann, U.R.; Papendieck, A.; Dahlems, U.; Gellissen, G.; Hoylaerts, M.F. Production and characterization of saratin, an inhibitor of von Willebrand factor-dependent platelet adhesion to collagen. *Semin Thromb Hemost* **2001**, *27*, 337-348.
114. Huizinga, E.G.; Schouten, A.; Connoly, T.M.; Kroon, J.; Sixma, J.J.; Gros, P. The structure of leech anti-platelet protein, an inhibitor of haemostasis. *Acta Crystallogr D Biol Crystallogr* **2001**, *57*, 1071-1078.
115. Pérez, L.A.A.; Tabachnick, W.J. Apyrase activity and adenosine diphosphate induced platelet aggregation inhibition by the salivary gland proteins of *Culicoides variipennis*, the North American vector of bluetongue viruses. *Vet Parasitol* **1996**, *61*, 327-338.
116. Rigbi, M.; Orevi, M.; Eldor, A. Platelet aggregation and coagulation inhibitors in leech saliva and their roles in leech therapy. *Semin Thromb Hemost* **1996**, *22*, 273-278.
117. Agoncillo, A. Meta-analysis on the safety and efficacy of lumbrokinase in peripheral arterial disease. *Eur Heart J Acute Ca* **2021**, 10, zuab020.219.
118. Mihara, H.; Sumi, H.; Yoneta, T.; Mizumoto, H.; Ikeda, R.; Seiki, M.; Maruyama, M. A novel fibrinolytic enzyme extracted from the earthworm, *Lumbricus rubellus*. *Jpn J Physiol* **1991**, *41*, 461-472.
119. Iannucci, N.B.; Camperi, S.A.; Cascone, O. Purification of lumbrokinase from *Eisenia fetida* using aqueous two-phase systems and anion-exchange chromatography. *Sep Purif Technol* **2008**, *64*, 131-134.
120. Bobrovsky, P.; Manuvera, V.; Baskova, I.; Nemirova, S.; Medvedev, A.; Lazarev, V. Recombinant destabilase from *Hirudo medicinalis* is able to dissolve human blood clots In vitro. *Curr Issues Mol Biol* **2021**, *43*, 2068-2081.

121. Baskova, I.P.; Nikonov, G.I. Destabilase: an enzyme of medicinal leech salivary gland secretion hydrolyzes the isopeptide bonds in stabilized fibrin. *Biokhimiia* **1985**, *50*, 424-431.
122. Kurdyumov, A.S.; Manuvera, V.A.; Baskova, I.P.; Lazarev, V.N. A comparison of the enzymatic properties of three recombinant isoforms of thrombolytic and antibacterial protein--Destabilase-Lysozyme from medicinal leech. *BMC Biochem* **2015**, *16*, 27.
123. Castellano, I.; Merlino, A. Gamma-glutamyl transpeptidases: Structure and function. In: Gamma-glutamyl transpeptidases, Springer, Basel, 2013.
124. Friedrich, T.; Kröger, B.; Koerwer, W.; Strube, K.H.; Meyer, T.; Bialojan, S. An isopeptide bond splitting enzyme from *Hirudo medicinalis* similar to gamma-glutamyl transpeptidase. *Eur J Biochem* **1998**, *256*, 297-302.
125. Reverter, D.; Vendrell, J.; Canals, F.; Horstmann, J.; Avilés, F.X.; Fritz, H.; Sommerhoff, C.P. A carboxypeptidase inhibitor from the medical leech *Hirudo medicinalis*. Isolation, sequence analysis, cDNA cloning, recombinant expression, and characterization. *J Biol Chem* **1998**, *273*, 32927-32933.
126. Arolas, J.L.; Castillo, V.; Bronsoms, S.; Aviles, F.X.; Ventura, S. Designing out disulfide bonds of leech carboxypeptidase inhibitor: implications for its folding, stability and function. *J Mol Biol* **2009**, *392*, 529-546.
127. Jung, H. Hyaluronidase: An overview of its properties, applications, and side effects. *Arch Plast Surg* **2020**, *47*, 297-300.
128. Linker, A.; Hoffman, P.; Meyer, K. The hyaluronidase of the leech: an endoglucuronidase. *Nature* **1957**, *180*, 810-811.
129. Das, B.K. An overview on hirudotherapy/leech therapy. *Ind ResJ Pharm Sci* **2014**, *1*, 33-45.
130. Hovingh, P.; Linker, A. Hyaluronidase activity in leeches (Hirudinea). *Comp Biochem Physiol B Biochem Mol Biol* **1999**, *124*, 319-326.
131. Feyertag, F.; Alvarez-Ponce, D. Disulfide bonds enable accelerated protein evolution. *Mol Biol Evol* **2017**, *34*, 1833-1837.
132. Andrade, M.A.; Perez-Iratxeta, C.; Ponting, C.P. Protein repeats: structures, functions, and evolution. *J Struct Biol* **2001**, *134*, 117-131.
133. Salzet, M. Anticoagulants and inhibitors of platelet aggregation derived from leeches. *FEBS Lett* **2001**, *492*, 187-192.
134. Electricwala, A.; Sawyer, R.T.; Jones, C.P.; Atkinson, T. Isolation of thrombin inhibitor from the leech *Hirudinaria manillensis*. *Blood Coagul Fibrinolysis* **1991**, *2*, 83-89.
135. Markwardt, F. Hirudin as alternative anticoagulant—a historical review. *Semin Thromb Hemost* **2002**, *28*, 405-414.