Single-Molecule Long-Read Sequencing Facilitates Transcriptomic Research For *Whitmania Pigra*, The Most Utilized Medical Leech In Chinese Traditional Medicine For Anticoagulant Therapy

Jing Song
Shaanxi Normal University
Ping Li
Shaanxi Normal University
De-Long Guan
Shaanxi Normal University
Yan Sun (sunyan@snnu.edu.cn)
Shaanxi Normal University

Research Article

Keywords: Whitmania pigra, leech, PacBio isoform sequencing, gene functions, genetic background

Posted Date: June 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-558486/v1

Abstract

Although leeches are of great medical and economic value in anticoagulant therapy, full-length transcriptomes for leeches remain scarce. Here, we generated the first full-length transcriptome for the paddy leech *Whitmania pigra* (the most widely utilized medical leech in Chinese traditional medicine) through Pacific Biosciences (Pacbio) single-molecule long-read sequencing. A total of 191,676 full-length non-chimeric (FLNC) reads were obtained, 30,660 were high-quality unique full-length transcripts. The BUSCO (Bench-marking Universal Single-Copy Orthologues) accession of completeness demonstrated that 74.8% of BUSCOs were complete. We functionally annotated 28,144 transcripts were in public databases, including NR, gene ontology (GO), Pfam, etc. Furthermore, 1,314 long non-coding RNAs (LncRNAs), 2,574 alternative splicing (AS) events, 932 transcript factors (TFs), and 33,258 simple sequence repeats (SSRs) we identified across all transcripts. From the generated data, a total of 426 anticoagulant genes, including 122 Antistasins, 124 with the Fibrinogen beta and gamma chains, and 62 Kazal-type serine protease inhibitors were screened out. Twenty-five novel proteins were revealed following the evaluation of the annotations and products of these anticoagulant transcripts. The regulation network between LncRNAs and corresponding coding transcripts was found with the typical mang-to-many pattern, especially obvious in a specific type of protein, Guamerin. Collectively, the present findings provide a rich set of full-length cDNA sequences for *W. pigra*, which will greatly facilitate research on transcriptomic genetic for this species and leeches.

Introduction

The paddy leech, *Whitmania pigra* (Whitman, 1884), is a species widely distributed in local freshwater ecosystems and is native to East Asia. (Shen et al., 2011; Kuo and Lai, 2019). Reports suggest that *W. pigra* is microphagous, non-blood-sucking and mostly prey on field snails (Weisblat, 2003; Kuo and Lai, 2019). Although it exhibits nonblood feeding character, many researchers have identified anticoagulation peptides from this leech and located the molecular basis to the synthesis of such substances (Chu et al., 2016; Khan et al., 2019a; Khan et al., 2019b; Hu et al., 2020; Huang et al., 2020). Phylogenetically, *W. pigra* belongs to the family Hirudinidae. It shares genome-wide orthologues genes with several blood-feeding species, including the medicinal leech *Hirudo medicinalis* and *Hirudo nipponia* (Kvist et al., 2013; Kvist et al., 2020). Several anticoagulant genes have been cloned or assembled successfully using various biological and chemical strategies to yield valuable contents including Antistasin, Guamerin, and Bdellastasin (Chu et al., 2016; Liu et al., 2016; Ren et al., 2016; Ren et al., 2019; Wang et al., 2019).

W. pigra has been treated as a medical leech in traditional Chinese medicine because it can effectively synthesize anticoagulant substances(Zhang et al., 2013). Other than blood-sucking leeches, the whole body of *W. pigra* can be administered medically rather than secretions, but show mild and controllable effects. With these advantages, *W. pigra* remains the most utilized medical leech in China. Its products have been applied in clinics to promote blood circulation and relieve vascular congestions (Zhong et al., 2008; Zhang et al., 2013; Ren et al., 2019; Wang et al., 2019; Hu et al., 2020). Notably, *W. pigra* has the ket for anticoagulant therapy. Therefore, it has a

broad commercial prospect and remains the focus of scientific research aimed at elucidating its pharmaceutical value.

Numerous studies have targeted the molecular foundation of anticoagulants in *W. pigra*, including several previous short-read transcriptomes. However, analyses lack high qualified full-length transcript sequences, labeling them ineffective and inaccurate. The reads produced by short-read sequencing were retrieved from fragmented cDNA libraries, and require assembly processes for reconstruction into transcripts. This will inevitably introduce misinterpretations to different isoforms and may merge different members of multigene families. Also, the short-read transcriptomes could not resolve important factors such as isoforms, long-non-coding RNAs (lncRNA), and alternative splicing (AS) events, which limits the understanding of the deeper meaning of related life activities (Rhoads and Au, 2015; Weirather et al., 2017; Hong et al., 2020). The anticoagulant repertoire obtained from these studies is relatively unitary, leaving room for further improvement of accuracy.

Due to the rapid advancement in long-read sequencing technologies, for example, Pacbio sequencing platforms, full-length and high-integrated transcriptomics has proven to be a more powerful tool for functional genetic studies on various types of organisms (Morin et al., 2018). For instance, long-read transcriptomics no longer requires an assembly process, and only a few processing steps are needed to generate accurate and full-length sequences ranging from $5 \star ts \rightarrow 3$ ends. This enhances the characterization of different types of isoforms and the resolution of the complexity of AS events. Besides protein-coding genes, these transcriptomes can also reveal different types of genetic motifs, such as transcription factors (TFs), simple sequence repeats (SSRs), and long noncoding RNAs (lncRNAs), which are valuable in the exploration of interesting regulation mechanisms (Rhoads and Au, 2015; Weirather et al., 2017).

Herein, we report a full-length transcriptome of *W. pigra* for further functional and genetic studies using the Pacbio isoform sequencing technique (Camacho et al., 2015; Yuan et al., 2019). The dataset, both the protein-coding genes and various genetic features reported in this work, is the first accurate transcriptomic genetic background for *W. pigra*. The findings provide an updated anticoagulant gene repertoire and shed light on its regulation mechanisms, and are expected to facilitate the application of *W. pigra* and other leeches species.

Results

Full-length transcriptome for *W. pigra*

We generated the full-length transcriptome dataset of *W. pigra* using a pooled sample via the Pacbio platform. Overall, 207,215 circular consensus sequences (CCS) of 705.78 Mb, were retrieved. These CCSs were clustered and assembled into 191,676 (92.50%) non-redundant full-length non-chimeric reads (FLNC). All FLNCs (multiple copies of the same transcript) were then clustered. For each obtained cluster, a consensus isoform was generated, a total of 53,467 high-quality (HQ) isoforms with accuracy greater

redundant sequences using the CD-HIT

software and obtained 30,660 transcripts (a total length of 93.19 Mb). Table 1 outlines the statistical summary for lengths and sizes of CCS, FLNC, and HQ isoforms and transcripts. The length distributions of these sequences depict their consistency and quality (Supplement Fig. 1 ~ 4). In particular, the connectivity of transcripts was much higher than that of previously assembled transcriptomes (all are averagely below 1000 bp long) and is comparable to the high-quality genome of another medical leech, *Hirudinaria manillensis* (average length of 2963.1 bp). The transcripts served as the final molecular sequence pool for screening gene components and genetic motifs. Based on BUSCO (Bench-marking Universal Single-Copy Orthologues) completeness analysis, 74.8% of BUSCOs were complete, among which 25.8% were duplicates (Fig. 1A).

Table 1
Sequence statistics for full-length transcriptome of *W.pigra*.

File type	Nucleotides (bp)	Sequences	GC (%)	Min Sequence Length (bp)	Max Sequence Length (bp)	Mean length (bp)	Total Size (Kb)
CCS	705,787,087	207,215	44.30	201	17590	3406.1	175,433
FLNC	619,526,806	191,676	46.01	50	17506	3232.2	154,138
HQ isoforms	159,075,749	53,467	43.30	50	17506	2975.2	39,733
Transcripts	93,199,769	30,660	42.40	52	17506	3039.8	23,348

All HQ transcripts were annotated in public databases, including NT, NR, Pfam GO, COG/KOG, KEGG, Swiss-Prot, and eggnog (Table S1). We annotated 28,144 transcripts accounting for 91.79% of the total number. More transcripts were annotated in NR (27,737), followed by eggnog (25,433), and Pfam (24,538). The homologous species distribution of *W. pigra* annotated in the NR database demonstrated that its gene sequences exhibited high homology with several other Annelids, including the freshwater leech *Helodbella robusta* (16171, 58.38%), and the sea worm *Capitella teleta* (2401, 8.67%) (Fig. 1B). GO classifications assigned 13,723 transcripts into 50 specific GO terms. The categories with the largest number of transcripts in each classification were as follows: Molecular function (binding, 6,523), biological process (cellular process, 5,788), and cellular component (cell part, 4,384) (Supplement Fig. 5).

We identified 24,719 CDSs from all transcripts, with a total length of 36.47 Mb. Through a combination of four different prediction methods, the Coding-Non-Coding Index (CNCI) (Sun et al., 2013), the Coding Potential Calculator (CPC) (Kong et al., 2007), the Pfam-scan (Pfam) (Finn et al., 2016) and the Coding Potential Assessment Tool (CPAT) (Li et al., 2014), 1,314 LncRNAs were revealed (Fig. 1C). We successfully located the corresponding target CDSs for nearly all LncRNAs (1,308 out of 1,314). The total number of targets was 12,775. In total, 2,574 alternative splicing (AS) events were predicted, and 338 of these transcripts were subjected to several origins more than once. Also, 932 transcript factors (TFs) were predicted. The zf-C2H2 was the most abundant TF. In addition, 33,258 SSRs were found, which comprised 5,592 mononucleotides, 1,485 dinucleotides, 24,737 trinucleotides, 1,162 tetranucleotides, 100

pentanucleotides, and 182 hexanucleotides. The number of SSR-containing sequences was 11,478. All the data retrieved in this work, including the sequences of transcripts, along with the annotation information, genetic motifs (LncRNAs, SSR, TFs), predicted alternative splicing, and their target CDSs, served as solid genetic database for W.pigra, and were deposited in Zenodo using the link: https://doi.org/10.5281/zenodo.4707300.

Identification of anticoagulant genes

We searched for all known types of anticoagulant genes that have been reported in leeches from the Swiss-prot database according to the annotations of transcripts. A total of 426 transcripts met such criteria (Supplementary File S1). These transcripts were further assigned into 8 Pfam protein families, including the antistasin family (121 genes, PF02822.13), Destabilase (14 genes, PF05497.11), Kazal-type serpin (62 genes, PF07648.14), Lectin C (72 genes, PF00059.20), Fibrinogen beta and gamma chains C-terminal globular domain (124 genes, PF00147.17), Eglin C (12 genes, PF00280.17), Cystatin family (17 genes, PF00031.20) and Neurohemerythrin (5 genes, PF01814.22). Among them, the transcript sequences for 16 genes were novel, yielded by alternative-splicing events, including a member of the Antistasin family (JXZ_transcript_6437) exhibiting a direct anticoagulant effect. Its nucleotide sequence had a gap when compared to the origin sequence (JXZ_transcript_50726) but encoded the same protein (Supplement Figure S6).

Meanwhile, we found 25 transcripts with no blast hits in the NR and swiss-prot databases (Supplement Table S2). They were suggested to produce newly identified anticoagulant proteins, which complement the existing anticoagulant reservoir of leeches and have proven that although it has been studied for years, *W. pigra* still has the potential on digging functional components. The new anticoagulant proteins include 6 antistasins, 3 Kazal-type serpins, 13 Lectin C, and 3 Cystatins. The conserved domains for their translated proteins were illustrated using seqlogo maps to depict the composition and variation pattern of their sequences (Fig. 2). The seqlogo maps demonstrated that the proteins with Lectin C domains harbor the most complex variation in sequences, whereas members of the Antistasin family all comprise typical Cysteine-rich structures.

LncRNA regulation network for anticoagulant genes

LncRNAs are regulation factors for protein-coding sequences. In previous sections, we have predicted the target sequences for all 1,314 LncRNAs and revealed 12,775 targets. Their relationships will guide future genetic studies such as inhibition or activation of focused gene expressions through manipulation of specific LncRNAs. Based on the linkages between the LncRNAs and target transcripts, we have generated a linkage map for all 426 anticoagulant transcripts (Supplement Figure S7). Notably, we found a specific node linked to a single type of transcript, encoding the Guamerin protein. This node represents a regulation mechanism with highly abundant factors. We have illustrated its regulation network in Fig. 3. The elements of 3 LncRNAs (JXZ_transcripts_27429, JXZ_transcripts_39693, and JXZ_transcripts_15344) are shown with the typical many-to-many correspondence trend with the coding

ion mechanism.

Discussion

Previous studies have noted the importance of accurate molecular sequences in exploring the genetic foundation of biological characters. The advantages of long-read sequencing technology represented by PacBio have been extensively demonstrated by recent studies (Grabherr et al., 2011; Rhoads and Au, 2015; Byrne et al., 2019; Zhang et al., 2019; Hong et al., 2020). The long-read sequencing technology can enrich and/or improve the current molecular sequences for a given organism. The initial objective of the present work was to obtain the first full-length transcriptome of a medical leech, *W. pigra*, widely used in Chinese traditional medicine. Of interest, we have successfully generated a high-qualified transcript dataset to enrich the current genetic resource for *W. pigra* and leeches. According to the length distribution and completeness of BUSCO analysis of all obtained transcripts, we believe that our data provide the best connected and most accurate transcriptome in leeches (Liu et al., 2018a; Liu et al., 2018b; Lu et al., 2018; Northcutt et al., 2018; Iwama et al., 2019; Khan et al., 2019a; Babenko et al., 2020). The average length of transcripts and their corresponding proteins are comparable to draft genomes, including the genome of *H. manillensis* (Guan et al., 2019; Babenko et al., 2020; Kvist et al., 2020). Besides, the important information retrieved and predicted from these transcripts will contribute to future studies. For instance, the AS events and the association between the LncRNA and target coding sequences will fuel further studies on regulation mechanisms or gene expression regulation. The sites of SSRs will guide researchers to design polymorphic primers.

Moreover, to prove that our data is valuable and an improved molecular resource for genetic studies of *W. pigra* and leeches, we established the transcripts with anticoagulant effects. Overall, 426 focused anticoagulant transcripts were screened out, among which 25 produce novel proteins. Compared to previously described transcriptomes of leeches, our results have noticeably improved the completeness in sequences and the identified number is much larger (Ren et al., 2016; Lu et al., 2018; Byrne et al., 2019; Kuo and Lai, 2019; Liu et al., 2019; Iwama et al., 2021). Combined with other obtained facts: the types of their translated proteins are mostly conserved which belong to 8 conserved protein families; there is a many-to-many pattern for LncRNAs to control their expression. Such increase in the number of transcripts indicate that the advantages of our data represented in distinguishing various types of isoforms. These sequences will promote the future application of this species. Also, the accurate long nucleotide sequences are essential references in the exploration of the complexity of the associated biological synthesis mechanisms (Northcutt et al., 2018; Zhang et al., 2019; Hong et al., 2020). Generally, more isoforms are suggested to contribute to the diverse complexity of regulatory mechanisms. Of note, the present findings also reflash our understanding of the biological characteristics of *W. pigra* and leeches, that their bio-synthesis of anticoagulant substances and the synthesis of target proteins are far more complex than previously thought.

However, our study has a few limitations. Because whole genomes for *W. pigra* were lacking, we could not classify the types of AS events and establish the actual binding site and inter-actions of LncRNA and coding sequences. Whether the LncRNA activates or suppresses the expression of their targets remains

, the full-length transcriptome only represents

the limit set of expressed transcripts from a single sample, so it is still part of the whole genetic background.

In conclusion, the present study has provided the full-length transcriptome for *W. pigra*, a medical leech in Chinese traditional medicine. This data, to our knowledge, is the most complete molecular database for this species currently available. It, therefore, enriches the genetic resource and will facilitate the bio-synthesis applications of this species and leeches.

Materials And Methods

Sample collection

Adult *W. pigra* samples were collected from the natural populations in a crop field in Wuhu, Anhui Province on February 26, 2021. The mussel tissues of the collected samples were dissected and immediately immersed in liquid nitrogen and kept at -80°C in a freezer.

RNA extraction

Total RNA was extracted from the mussel tissues using the TRIzol reagent (Invitrogen, Carlsbad, CA, USA) following the manufacturer's protocol. RNA degradation and contamination were evaluated using 1% agarose gel electrophoresis. The integrity and purity of RNA were assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and the NanoDrop 2000 (Thermo Scientific, Wilmington, DE, USA).

Library preparation and sequencing

The Isoform Sequencing cDNA library was prepared according to the Isoform Sequencing protocol employing the Clontech SMARTer PCR cDNA Synthesis Kit and the BluePippin Size Selection System protocol (Pacific Biosciences (PN 100-092-800-03)). The pooled RNA samples were reverse transcribed and sequenced by Beijing Biomarker Biotechnology Company (Biomarker, Beijing, CN).

Data analysis

Sequence data were processed via the SMRT Link 5.1.0 software. CCSs were generated from subread BAM files with the CCS module. FLNC fasta files were generated and fed into the clustering step, which subsequently performed with isoform-level cluster and polish by ICE module and Arrow polish module. To obtain final transcripts, we eliminated any redundancy in corrected consensus reads using the CD-Hit software (Fu et al., 2012). The completeness of the transcriptome was assessed using the BUSCO v3.0.2 software with the arthropoda_odb9 database (number of BUSCOs: 1066) (Seppey et al., 2019).

Gene functional annotation

Gene function was annotated based on the following public databases: Non-redundant protein sequences (NR) (Li et al., 2002), non-redundant nucleotide sequences (NT), protein family (Pfam),

clusters of orthologous groups of proteins (KOG) (Tatusov et al., 2003), Swiss-Prot (Bairoch and Apweiler, 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2004), and gene ontology (GO) (Ashburner et al., 2000), and evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggnog) (Huerta-Cepas et al., 2019). The basic local alignment search tool (BLAST) was employed for Nt/Nr and Swiss-Prot database analyses, whereas the Diamond BLASTX software was applied with an e-value set to '1e-10' in the COG/KOG and KEGG database analyses. The Hmmerscan (El-Gebali et al., 2019) and Pfam2GO (https://github.com/am8265/Pfam2GO) software were used in Pfam and GO database analyses. All the software were set to default parameters.

CDS prediction

The CDSs from cDNAs were established via the ANGEL (Shimizu et al., 2006) pipeline. Firstly, we adopted the TransDecoder v3.0.1 (Haas et al., 2013) to identify CDS, followed by a translation into confident proteins. These confident protein sequences were utilized for ANGEL training and prediction of the complete sequence set.

Determining genetic motifs

The CNCI (Sun et al., 2013), CPC (Kong et al., 2007), Pfam (Finn et al., 2016), and CPAT (Li et al., 2014) methods were employed to predict the coding potential of transcripts. Then, transcripts via all the above four tools were filtered out. Those without coding potential acted as the candidate set of LncRNAs. The TFs were predicted via the animalTFDB 2.0 database (Zhang et al., 2015). SSRs of the transcriptome were identified using MIcroSAtellite v1.0 (MISA), which allowed for the identification and localization of perfect microsatellites, and compound microsatellites that were interrupted by a certain number of bases.

Determination of anticoagulants and sequence analysis

We filtered out transcripts related to anticoagulant effects according to their annotations. All known anticoagulant proteins available in the swiss-prot database were selected and used as query sequences. To extract and process the sequences, Geneious software was employed. Alignment of these sequences and their proteins was performed using the MAFFT (Katoh and Standley, 2013) software and illustrated in Geneious (Kearse et al., 2012) software.

Declarations

Data availability statement

The datasets presented in this study are available in online repositories. The names of the repository/repositories and accession number(s) can be found at zenodo; https://doi.org/10.5281/zenodo.4707300.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

J.S and P.L conceived the study and designed the experiments. J.S performed the sequencing experiments. D-LG analyzed the data. J.S and D-L.G wrote the manuscript. J.S and Y.S revised the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Excellent Doctor Innovation Project of Shaanxi Normal University (S2015YB03), Fundamental Research Funds for the Central Universities (GK201903063).

Acknowledgments

We thank the Beijing Biomarker Biology Company (Beijing, China) for technical assistance with sequencing. We also thank Home for Researchers editorial team for their scientific editing.

References

1. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1), 25-29. doi: 10.1038/75556.
2. Babenko, V.V., Podgorny, O.V., Manuvera, V.A., Kasianov, A.S., Manolov, A.I., Grafskaia, E.N., et al. (2020). Draft genome sequences of Hirudo medicinalis and salivary transcriptome of three closely related medicinal leeches. *BMC Genomics* 21(1), 331. doi: 10.1186/s12864-020-6748-0.
3. Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(1), 45-48. doi: 10.1093/nar/28.1.45.
4. Byrne, A., Cole, C., Volden, R., and Vollmers, C. (2019). Realizing the potential of full-length transcriptome sequencing. *Philos Trans R Soc Lond B Biol Sci* 374(1786), 20190097. doi: 10.1098/rstb.2019.0097.
5. Camacho, J.P., Ruiz-Ruano, F.J., Martín-Blázquez, R., López-León, M.D., Cabrero, J., Lorite, P., et al. (2015). A step to the gigantic genome of the desert locust: chromosome sizes and repeated DNAs. *Chromosoma* 124(2), 263-275. doi: 10.1007/s00412-014-0499-0.
6. Chu, F., Wang, X., Sun, Q., Liang, H., Wang, S., An, D., et al. (2016). Purification and characterization of a novel fibrinolytic enzyme from Whitmania pigra Whitman. *Clin Exp Hypertens* 38(7), 594-601. doi: 10.3109/10641963.2016.1174254.

7. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res* 47(D1), D427-d432. doi: 10.1093/nar/gky995.
8. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23), 3150-3152. doi: 10.1093/bioinformatics/bts565.
9. Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7), 644-652. doi: 10.1038/nbt.1883.
10. Guan, D.L., Yang, J., Liu, Y.K., Li, Y., Mi, D., Ma, L.B., et al. (2019). Draft Genome of the Asian Buffalo Leech Hirudinaria manillensis. *Front Genet* 10, 1321. doi: 10.3389/fgene.2019.01321.
11. Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., et al. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8), 1494-1512. doi: 10.1038/nprot.2013.084.
12. Hong, F., Mo, S.H., Lin, X.Y., Niu, J., Yin, J., and Wei, D. (2020). The PacBio Full-Length Transcriptome of the Tea Aphid as a Reference Resource. *Front Genet* 11, 558394. doi: 10.3389/fgene.2020.558394.
13. Hu, B., Xu, L., Li, Y., Bai, X., Xing, M., Cao, Q., et al. (2020). A peptide inhibitor of macrophage migration in atherosclerosis purified from the leech Whitmania pigra. *J Ethnopharmacol* 254, 112723. doi: 10.1016/j.jep.2020.112723.
14. Huang, Q., Gao, Q., Chai, X., Ren, W., Zhang, G., Kong, Y., et al. (2020). A novel thrombin inhibitory peptide discovered from leech using affinity chromatography combined with ultra-high performance liquid chromatography-high resolution mass spectroscopy. *J Chromatogr B Analyt Technol Biomed Life Sci* 1151, 122153. doi: 10.1016/j.jchromb.2020.122153.
15. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K., Cook, H., et al. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 47(D1), D309-d314. doi: 10.1093/nar/gky1085.
16. Iwama, R., Oceguera-Figueroa, A., Giribet, G., and Kvist, S. (2019). The salivary transcriptome of Limnobdella mexicana (Annelida: Clitellata: Praobdellidae) and orthology determination of major leech anticoagulants. *Parasitology* 146(10), 1338-1346. doi: 10.1017/s0031182019000593.
17. Iwama, R.E., Tessler, M., Siddall, M.E., and Kvist, S. (2021). The Origin and Evolution of Antistasin-like Proteins in Leeches (Hirudinida, Clitellata). *Genome Biol Evol* 13(1). doi: 10.1093/gbe/evaa242.
18. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32(Database issue), D277-280. doi: 10.1093/nar/gkh063.
19. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4), 772-780. doi: 10.1093/molbev/mst010.
20. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of [illegible] doi: 10.1093/bioinformatics/bts199.

21. Khan, M.S., Guan, D.L., Kvist, S., Ma, L.B., Xie, J.Y., and Xu, S.Q. (2019a). Transcriptomics and differential gene expression in Whitmania pigra (Annelida: Clitellata: Hirudinida: Hirudinidae): Contrasting feeding and fasting modes. *Ecol Evol* 9(8), 4706-4719. doi: 10.1002/ece3.5074.
22. Khan, M.S., Guan, D.L., Ma, L.B., Xie, J.Y., and Xu, S.Q. (2019b). Analysis of synonymous codon usage pattern of genes in unique non-blood-sucking leech Whitmania pigra. *J Cell Biochem* 120(6), 9850-9858. doi: 10.1002/jcb.28267.
23. Kuo, D.H., and Lai, Y.T. (2019). On the origin of leeches by evolution of development. *Dev Growth Differ* 61(1), 43-57. doi: 10.1111/dgd.12573.
24. Kvist, S., Manzano-Marín, A., de Carle, D., Trontelj, P., and Siddall, M.E. (2020). Draft genome of the European medicinal leech Hirudo medicinalis (Annelida, Clitellata, Hirudiniformes) with emphasis on anticoagulants. *Sci Rep* 10(1), 9885. doi: 10.1038/s41598-020-66749-5.
25. Kvist, S., Min, G.S., and Siddall, M.E. (2013). Diversity and selective pressures of anticoagulants in three medicinal leeches (Hirudinida: Hirudinidae, Macrobdellidae). *Ecol Evol* 3(4), 918-933. doi: 10.1002/ece3.480.
26. Li, W., Jaroszewski, L., and Godzik, A. (2002). Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics* 18(1), 77-82. doi: 10.1093/bioinformatics/18.1.77.
27. Liu, X., Wang, C., Ding, X., Liu, X., Li, Q., and Kong, Y. (2016). A novel selective inhibitor to thrombin-induced platelet aggregation purified from the leech Whitmania pigra. *Biochem Biophys Res Commun* 473(1), 349-354. doi: 10.1016/j.bbrc.2016.03.117.
28. Liu, Z., Tong, X., Su, Y., Wang, D., Du, X., Zhao, F., et al. (2019). In-depth profiles of bioactive large molecules in saliva secretions of leeches determined by combining salivary gland proteome and transcriptome data. *J Proteomics* 200, 153-160. doi: 10.1016/j.jprot.2019.03.009.
29. Liu, Z., Wang, Y., Tong, X., Su, Y., Yang, L., Wang, D., et al. (2018a). De novo assembly and comparative transcriptome characterization of Poecilobdella javanica provide insight into blood feeding of medicinal leeches. *Mol Omics* 14(5), 352-361. doi: 10.1039/c8mo00098k.
30. Liu, Z., Zhao, F., Tong, X., Liu, K., Wang, B., Yang, L., et al. (2018b). Comparative transcriptomic analysis reveals the mechanism of leech environmental adaptation. *Gene* 664, 70-77. doi: 10.1016/j.gene.2018.04.063.
31. Lu, Z., Shi, P., You, H., Liu, Y., and Chen, S. (2018). Transcriptomic analysis of the salivary gland of medicinal leech Hirudo nipponia. *PLoS One* 13(10), e0205875. doi: 10.1371/journal.pone.0205875.
32. Morin, P.A., Foote, A.D., Hill, C.M., Simon-Bouhet, B., Lang, A.R., and Louis, M. (2018). SNP Discovery from Single and Multiplex Genome Assemblies of Non-model Organisms. *Methods Mol Biol* 1712, 113-144. doi: 10.1007/978-1-4939-7514-3_9.
33. Northcutt, A.J., Fischer, E.K., Puhl, J.G., Mesce, K.A., and Schulz, D.J. (2018). An annotated CNS transcriptome of the medicinal leech, Hirudo verbana: De novo sequencing to characterize genes associated with nervous system activity. *PLoS One* 13(7), e0201206. doi:

34. Ren, S.H., Liu, Z.J., Cao, Y., Hua, Y., Chen, C., Guo, W., et al. (2019). A novel protease-activated receptor 1 inhibitor from the leech Whitmania pigra. *Chin J Nat Med* 17(8), 591-599. doi: 10.1016/s1875-5364(19)30061-5.
35. Ren, Y., Yang, Y., Wu, W., Zhang, M., Wu, H., and Li, X. (2016). Identification and characterization of novel anticoagulant peptide with thrombolytic effect and nutrient oligopeptides with high branched chain amino acid from Whitmania pigra protein. *Amino Acids* 48(11), 2657-2670. doi: 10.1007/s00726-016-2299-8.
36. Rhoads, A., and Au, K.F. (2015). PacBio Sequencing and Its Applications. *Genomics Proteomics Bioinformatics* 13(5), 278-289. doi: 10.1016/j.gpb.2015.08.002.
37. Seppey, M., Manni, M., and Zdobnov, E.M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* 1962, 227-245. doi: 10.1007/978-1-4939-9173-0_14.
38. Shen, X., Wu, Z., Sun, M., Ren, J., and Liu, B. (2011). The complete mitochondrial genome sequence of Whitmania pigra (Annelida, Hirudinea): the first representative from the class Hirudinea. *Comp Biochem Physiol Part D Genomics Proteomics* 6(2), 133-138. doi: 10.1016/j.cbd.2010.12.001.
39. Shimizu, K., Adachi, J., and Muraoka, Y. (2006). ANGLE: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *J Bioinform Comput Biol* 4(3), 649-664. doi: 10.1142/s0219720006002260.
40. Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41. doi: 10.1186/1471-2105-4-41.
41. Wang, X., Niu, M., Wu, S.N., Hu, H.W., Liu, X.Y., Ma, S.Y., et al. (2019). Leeches attenuate blood hyperviscosity and related metabolic disorders in rats differently than aspirin. *J Ethnopharmacol* 238, 111813. doi: 10.1016/j.jep.2019.03.040.
42. Weirather, J.L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.J., et al. (2017). Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Res* 6, 100. doi: 10.12688/f1000research.10571.2.
43. Weisblat, D.A. (2003). Leeches. *Curr Biol* 13(19), R752. doi: 10.1016/j.cub.2003.09.011.
44. Yuan, H., Chang, H., Zhao, L., Yang, C., and Huang, Y. (2019). Sex- and tissue-specific transcriptome analyses and expression profiling of olfactory-related genes in Ceracris nigricornis Walker (Orthoptera: Acrididae). *BMC Genomics* 20(1), 808. doi: 10.1186/s12864-019-6208-x.
45. Zhang, H.M., Liu, T., Liu, C.J., Song, S., Zhang, X., Liu, W., et al. (2015). AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. *Nucleic Acids Res* 43(Database issue), D76-81. doi: 10.1093/nar/gku887.
46. Zhang, W., Zhang, R.X., Li, J., Liang, F., and Qian, Z.Z. (2013). Species study on Chinese medicine leech and discussion on its resource sustainable utilization. *Zhongguo Zhong Yao Za Zhi* 38(6), 914-918.
47. Zhang, X., Li, G., Jiang, H., Li, L., Ma, J., Li, H., et al. (2019). Full-length transcriptome analysis of olved in the innate immune system. *Fish*

Shellfish Immunol 87, 346-359. doi: 10.1016/j.fsi.2019.01.023.

48. Zhong, S., Yang, D.P., and Cui, Z. (2008). Studies on anticoagulant constituents in dried Whitmania pigra. *Zhongguo Zhong Yao Za Zhi* 33(23), 2781-2784.

Figures

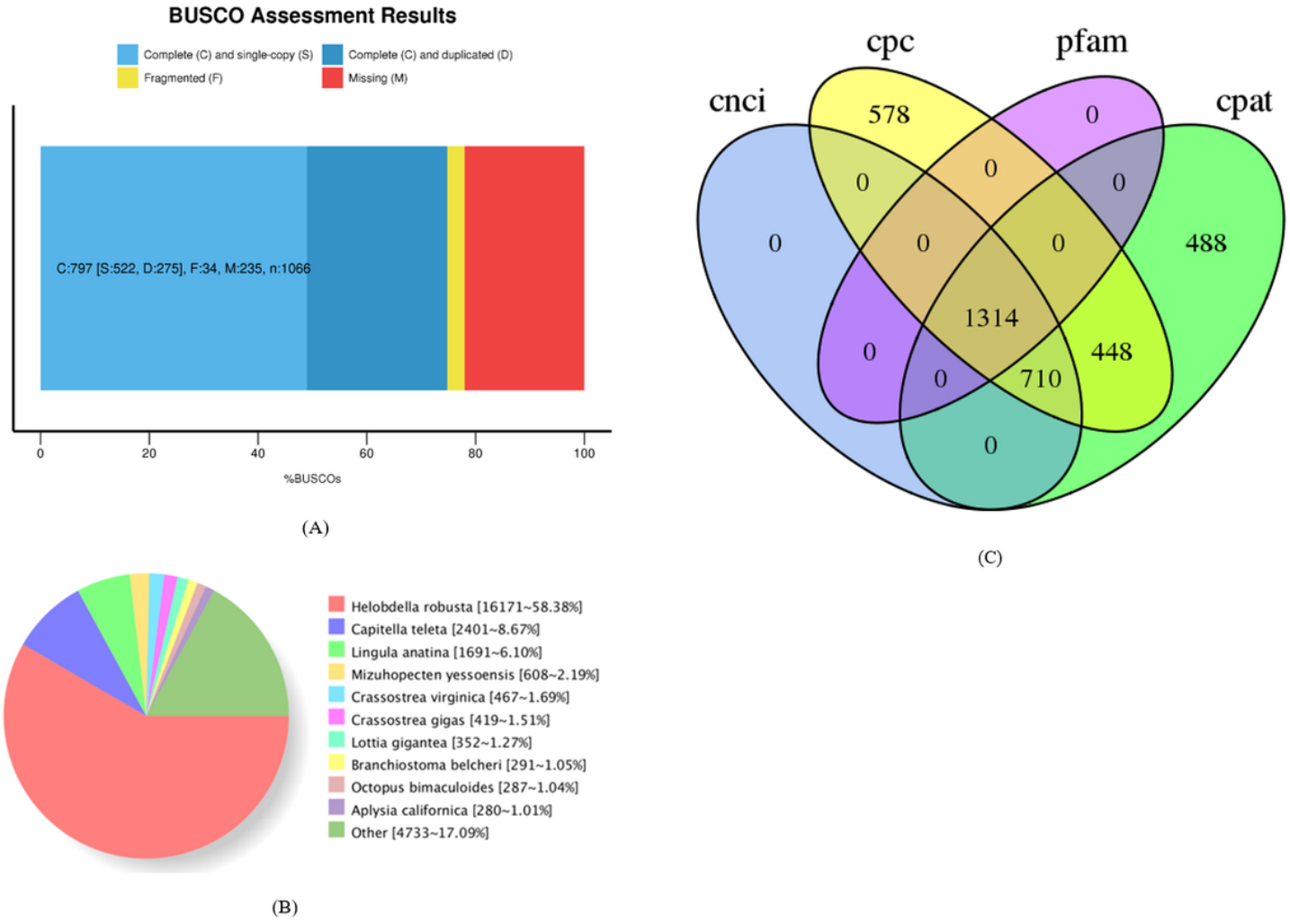

Figure 1

Basic descriptions of the generated full-length transcriptome of W. pigra. (A)BUSCO analysis of transcripts. The proportions are classified as complete (C, blues), complete duplicated (D, light blue), complete single-copy (S, dark blue), fragmented (F, Yellow), and missing (M, red). (B) Homologous species distribution of NR annotations in transcripts of W. pigra. (C) Long non-coding RNAs predicted by Calculator (CPC), Coding-Non-coding Index (CNCI), Coding Potential Assessment Tool (CPAT), and Pfam protein structure domain analysis.

Figure 2

Alignment of conserved domains for newly identified anticoagulant proteins. The consensus sequence and the seqlogo map for each alignment are shown above each alignment distribution. The colored boxes represent different functional areas; the red, orange, light blue, and green represent the domains of Antistasin, Kazal-type serpin, Cystatin, and Lectin C, respectively.

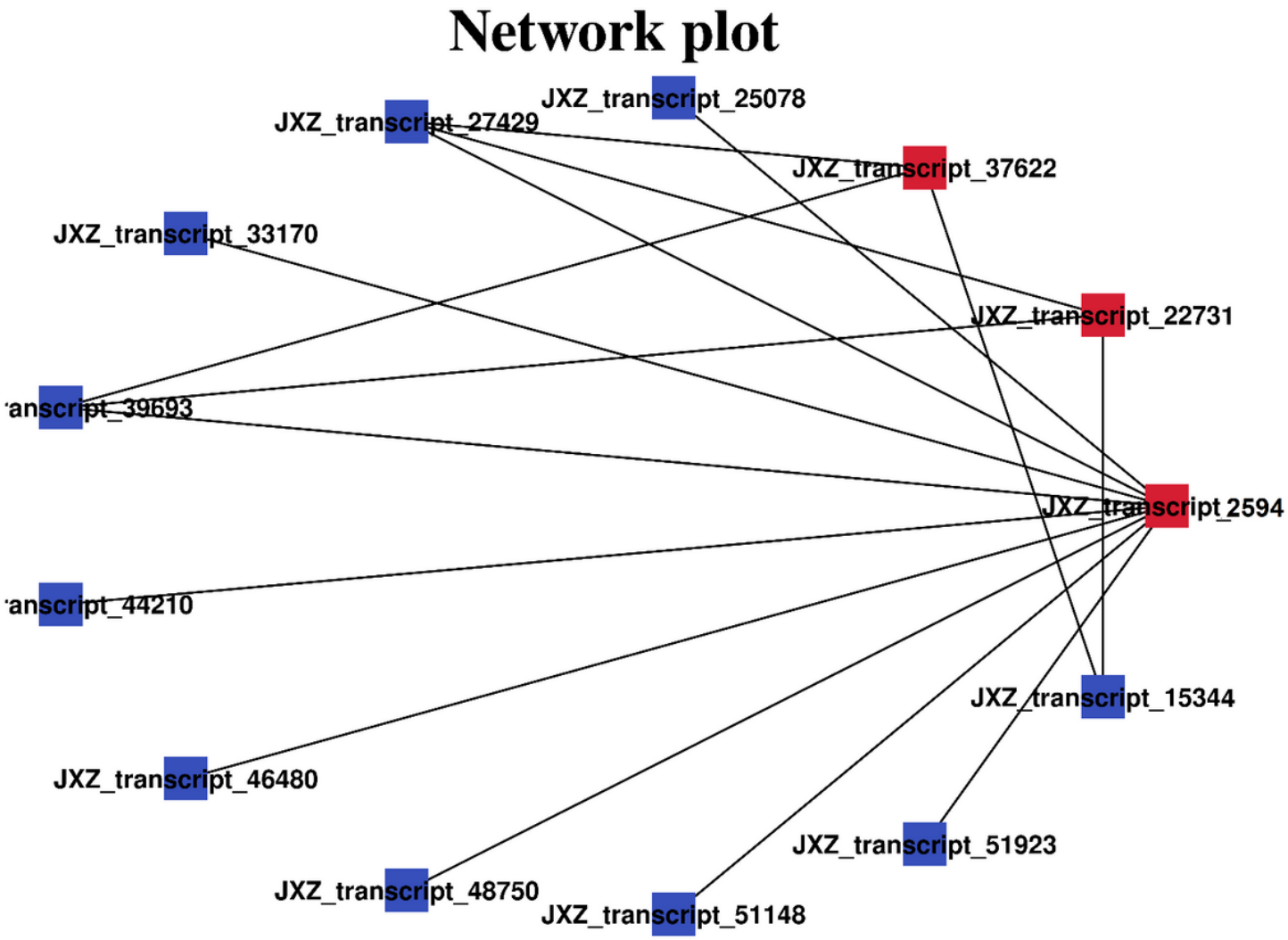

Figure 3

Regulation network of 3 isoforms that encode Guamerins. Red dots represent the 3 isoforms, whereas blue dots represent their corresponding LncRNAs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- FigureS1.jpg
- FigureS2.jpg
- FigureS3.jpg
- FigureS4.jpg
- FigureS5.png

- FigureS7.png
- TableS1.docx
- TableS2.docx