The Genome of Medicinal Leech (*Whitmania pigra*) and comparative genomic study for Exploration of Bioactive Ingredients

Lei Tong
Kunming University
Shao-Xing Dai
Kunming University of Science and Technology
De-Jun Kong
Kunming University
Peng-Peng Yang
Kunming University of Science and Technology
Xin Tong
Kunming University of Science and Technology
Xiang-Rong Tong
Kunming University
Xiao-Xu Bi
Kunming University
Yuan Su
Kunming University
Yu-Qi Zhao
University of California Los Angeles
Zi-Chao Liu (abclzc@aliyun.com)
Kunming University https://orcid.org/0000-0002-7509-6209

Research article

Keywords: Whitmania pigra, Genome, Bioactive Ingredients, Helobdella robusta, Hirudo medicinalis

Posted Date: October 12th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-31354/v2

Version of Record: A version of this preprint was published at BMC Genomics on January 24th, 2022. See the published version at https://doi.org/10.1186/s12864-022-08290-5.

Abstract

Background

Leeches are classic annelids that have a huge diversity and closely related to people, especially medicinal leeches. Medicinal leeches have been widely utilized in medicine based on the pharmacological activities of their bioactive ingredients. Comparative genomic study of these leeches enables us to understand the difference among medicinal leeches and other leeches and facilitates the discovery of bioactive ingredients.

Results

In this study, we reported the genome of *Whitmania pigra* and compared it with *Hirudo medicinalis* and *Helobdella robusta*. The assembled genome size of *W. pigra* is 177 Mbp, close to the estimated genome. Approximately about 23% of the genome was repetitive. A total of 26,743 protein-coding genes were subsequently predicted. *W. pigra* have 12346 (46%) and 10295 (38%) orthologous genes with *H. medicinalis* and *H. robusta*, respectively. About 20% and 24% genes in *W. pigra* showed syntenic arrangement with *H. medicinalis* and *H. robusta, respectively,* revealed by gene synteny analysis. Furthermore, *W. pigra, H. medicinalis* and *H. robusta* expanded different gene families enriched in different biological processes. By inspecting genome distribution and gene structure of hirudin, we identified a new hirudin gene g17108 (hirudin_2) with different cysteine pattern. Finally, we systematically explored and compared the active substances in the genomes of three leeches. The results showed that *W. pigra* and *H. medicinalis* exceed *H. robusta* in both kinds and gene number of active molecules.

Conclusions

This study reported the genome of *W. pigra* and compared it with other two leeches, which provides an important genome resource and new insight into the exploration and development of bioactive molecules of medicinal leeches.

Background

Leeches are segmented parasitic or predatory worms that belong to the phylum Annelida and the subclass Hirudinea with ability to extend or contract their bodies[1-3]. Most leeches live in freshwater environments, while some species can be found in terrestrial and marine environments. The best-known leeches, such as European medicinal leech *Hirudo medicinalis* are hematophagous, feeding on vertebrate blood and invertebrate hemolymph[4-6]. *H. medicinalis* attaches to the host by means of its two suckers and bites through the skin of its victim. Most leech species, however, are predatory, feeding primarily by swallowing other invertebrates. Almost 700 species of leeches are currently recognized, of which some 100 are marine species, 90 terrestrial and the remainder freshwater taxa.

Although a huge diversity and close relationship to people, we know little about the genome of leeches. In 2013, one leech *H. robusta* was sequenced to study bilaterian evolution [7]. *H. robusta* is a freshwater leech in the family glossiphoniidae, and a type of annelid with anterior and posterior suckers that are used for locomotion and feeding on blood. Its early development has been studied extensively. For another important family hirudinidae, the genome of *H. medicinalis* has just been published during the review period of this study[8, 9]. The family hirudinidae includes medicinal leeches which have been widely utilized in medical procedures for thousands of years. Because of their important bioactive ingredients, medicinal leeches, such as *H. medicinalis* and related species, have engendered great interest from pharmaceutical companies.

Comparative study of these available genomes of leeches facilitates the discovery of bioactive ingredients. In this study, we reported the genome of another medicinal leech *W. pigra* in the family hirudinidae and compared it with other two leeches (**Figure 1A**). *W. pigra*, an Asian freshwater leech, is non-blood feeding, despite the placement of this genus within the family hirudinidae [10]. The family hirudinidae also includes *H. medicinalis* and several other blood feeding species. *W. pigra* is considered macrophagous, suggesting that it commonly swallows or takes bites out of prey sources[11-13]. *W. pigra* is recorded in the current Chinese Pharmacopoeia as the source of leeches, and the most commonly available from Chinese commercial leech market[14]. We first analyzed the genome of *W. pigra* and conducted gene synteny analysis among the three leeches *H. robusta*, *W. pigra, and H. medicinalis.* Then we analyzed the expansion and contraction of gene family among seven related species (*H. robusta*, *Lottia gigantea*, *Capitella teleta*, *Schmidtea mediterranea*, *Schistosoma mansoni*, *W. pigra, H. medicinalis*). The sequence diversity, genome distribution and gene structure of hirudin were also studied. At last, we explored nine kinds of bioactive compounds in the genomes of the three leeches. This study pointed out the differences in the genome of the three leeches *W. pigra, H. medicinalis* and *H. robusta,* and provided insight into the exploration and development of the bioactive molecules of medicinal leeches.

Results

Summary of genome assembly and annotation for *W. pigra*

Using a whole-genome shotgun strategy with the Illumina HiSeq™ 2000 platform, we sequenced the genome of *W. pigra* from Wuhan, the provincial capital Hubei, China. The de novo assembly of a 146 Gbp high-quality sequences from 2 paired-end and 3 mate-pair libraries provided 100-fold coverage with a total assembly length of 177 Mbp (**Table 1**), which approximates the genome size estimated by 23 K-mer distribution (**Figure S1**). The scaffold N50 is 728 Kbp. 3495 scaffolds are with length >2 Kbp. Repeat content comprised 23% of the *W. pigra* genome, which is 10% lower than that of the *H. robusta* [7]. The *W. pigra* shares a similar profile of GC content (35%) with *H. robusta* (33%). A total of 26,743 protein-coding genes were predicted in *W. pigra*. *W. pigra* and *H. Robusta* showed similar gene model features in a whole. However, *W. pigra* has shorter intron length and longer protein length compared with *H. robusta* (Table 1). A total of 17123 protein-coding genes were annotated in all three common databases Uniprot,

TrEMBL and interPro (**Figure 1B**). We identified 12346 and 10295 orthologous genes between *W. pigra* and *H. medicinalis*, and between *W. pigra* and *H. robusta, respectively,* using reciprocal best blast hits (RBHs) method (**Figure 1C**). There are a large proportion of genes (14398 and 16449) in *W. pigra* not assigned as orthologous genes.

Table 1. Summary for genome sequencing, assembly and annotation

	H. robusta (Ref [7])	*W. pigra* (this study)	*H. medicinalis* (Ref [8])
Size of genome assembly	228 Mbp	177 Mbp	187M
Num. of Scaffolds	1,993	10,050	14,042
Num. of Scaffolds (> 2Kbp)	1124	3495	105
Scaffold N50	3,060 Kbp	728 Kbp	97 Kbp
Total reads	3,176,156	118,388,619	62,184,084
Reads mapping to genome (%)	2,839,951 (89%)	112,480,685 (95%)	NA
Sequencing coverage depth	7.92X	100X	73X
Repetitive content (%)	33	23	NA
GC (%)	33	35	41
Num. of predicted genes	23,400	26,743	14,596
Protein length (aa)	376	438	464
Mean exon length	203 bp	205 bp	224
Mean intron length	526 bp	391 bp	716
Mean number of exons per gene	6.1	6.4	8

GC, fraction of guanine plus cytosine nucleobases; **Scaffold N50**, the length such that half of the assembled sequence is in scaffolds longer than this length.

Gene synteny among the genomes of *W. pigra, H. medicinalis* and *H. robusta*

The above result showed that *W. pigra* only has 46.2% (12346) orthologous genes in *H. medicinalis, and* 38.4% (10295) orthologous genes in *H. robusta*. To further compare the genome similarity among the three leeches, we performed a careful analysis of syntenic blocks between *W. pigra and H. medicinalis,* and between *W. pigra and H. robusta* using MCScanX[15]. As small scaffolds are not useful for gene synteny analysis, we only considered the scaffold with more than 30 genes. Finally, we identified 21

scaffolds in *H. medicinalis* had syntenic blocks matched to the 13 scaffolds in *W. pigra.* In contrast, there are 33 scaffolds in *H. robusta* matched to the 21 scaffolds in *W. pigra* (**Figure 2**). Overall, the genome of *W. pigra* has a good collinearity relationship with other two genomes. We further examined the synthetic blocks in the larger scaffolds Wh8, wh9, wh17, and wh22. We found that *H. medicinalis* tends to have larger synthetic blocks matched to the scaffolds of *W. pigra* than *H. robusta.* It suggests that compared to *H. robusta, W. pigra* has a more similar genome structure with *H. medicinalis.*

The expansion and contraction of gene family in the *W. pigra* genome

After analysis of gene synteny, we further analyzed the expansion and contraction of gene family among the seven species: *H. robusta*, *Lottia gigantea*, *Capitella teleta*, *Schmidtea mediterranea*, *Schistosoma mansoni*, *W. pigra, H. medicinalis*. We compared the predicted proteomes of seven species, yielding a total of 13563 orthologous gene families that comprised 108245 genes. We found 1488, 832 and 1266 gene families expanded in *W. pigra, H. medicinalis* and *H. robusta*, respectively (**Figure 1D**). Of these families, there are 63, 1 and 59 families that are evolving rapidly (P<0.05) in *W. pigra, H. medicinalis* and *H. robusta*, respectively (**Figure 1E**). These rapidly evolving families are species-specific and little overlap between the two species (**Figure 1E**). To reveal the molecular function and structural domain of these rapidly evolving families, we performed enrichment analyses by gene ontology terms and interPro domains. The enrichment results showed a clear difference among the three leeches. For *W. pigra*, the expanded families are enriched in the following functions: protein histidine kinase activity, O−acyltransferase activity, thiamine pyrophosphate binding, carbohydrate binding, proteolysis and so on. For *H. robusta*, the expanded families are mainly enriched in the functions such as sodium channel activity, sodium ion transport, zinc ion binding, and RNA−DNA hybrid ribonuclease activity. For *H. medicinalis,* only two functions endopeptidase inhibitor activity and extracellular region are enriched (**Figure 3A**). In contrast, for the contracted families, there are little go terms enriched in *W. pigra* and *H. robusta, but more* go terms enriched *in H. medicinalis.* For example, iron ion binding, heme binding, proteolysis, and sodium channel activity functions are enriched by the contracted family *in H. medicinalis* (**Figure 3C**). Corresponding to these functions, specific protein domains are enriched in different leeches. These results imply the three species may take different adaptive strategies. And the different functions and domains are potentially related to environmental adaptation and bioactive peptides properties of the three leeches.

Phylogenetic analysis and sequence alignment of the hirudin gene family

As the most well-studied natural anticoagulant from leeches, hirudin has served as a standard for designing natural coagulation inhibitors[16]. Hirudin may be useful in the therapy of thrombosis because of its specific antithrombin effects[17]. We identified two hirudin genes g14352 and g17108 (**Figure 4A**) in *W. pigra* in this study. We named g14352 and g17108 as hirudin_1, hirudin_2, respectively (**Figure 5**). For comparison, we also identified three hirudin genes g9136, g9138, and g9139 in *H. medicinalis.* These five hirudin genes and 38 hirudin-like sequences from protein database UniProt were used to clarify the phylogenetic relationships of these hirudin genes (**Figure 4A**). They are clustered into three clades (named

Groups 1, 2 and 3) (**Figure 4A**). Three groups are highly supported with bootstrap value >95. The sequences (Group 3) from *W. pigra* do not cluster with the other hirudin genes. Groups 1, 2 and 3 follow different cysteine patterns CX(7)CX(1)CX(5)CX(5)CX(10)C, CX(7)CX(1)CX(5)CX(5)CX(8)C and CX(8)CX(1)CX(5)CX(5)CX(10)C, respectively (**Figure 4B**). The pattern of group 1 is the typical cysteine pattern of the hirudin. In contrast, gene g17108 (hirudin_2) of *W. pigra* shows the third cysteine pattern, which inserts an extra amino acid between the first and second cysteines. The gene g17108 (hirudin_2) is a new kind of hirudin, which has not been reported before.

Genome-wide distribution, gene structure and transcript levels of hirudin genes

Although there are a lot of studies about hirudin, the genome-wide distribution and gene structure of hirudin have not been reported. By sequence searching, we found that g14352 (hirudin_1) and g17108 (hirudin_2) are located at different scaffolds 5072 and 278, respectively (**Figure 5A**). The left and right sides of g17108 (hirudin_2) are surrounded by multiple genes (24 and 69 genes, respectively). We can infer that the two genes are separated by great distances (>210 Kbp). It suggested a lot of genome rearrangement events happened after gene duplication of hirudin genes. Furthermore, gene structures of the two hirudin genes are also different. g14352 (hirudin_1) only has three exons. In contrast, g17108 (hirudin_2) has four exons, which encode a signal peptide and a longer tail. Therefore, protein hirudin_2 has a longer sequence than hirudin_1 (**Figure 4B and 5B**).

Exploration of bioactive ingredients in the *W. pigra* genome

There are more than 20 bioactive substances identified from leeches, such as Antistasin, hirustasin, ghilantens, hirudin [2, 6, 18, 19]. These molecules have analgesic, anti-inflammatory, anticoagulant, platelet inhibitory, thrombin regulatory functions, and so on. *W. pigra, H. medicinalis* and *H. robusta* are belong to the family hirudinidae and glossiphoniidae, respectively. The detailed distribution of these bioactive substances in different leech species is still unknown. It is essential to identify and compare these active molecules in different leeches. Using the genome data, we systematic explored and compared five classes of active substances in *W. pigra, H. medicinalis* and *H. robusta* (**Table 2**). All 9 common active molecules were found in *W. pigra*. It is noteworthy that hirustasin, hirudin and destabilase I genes are absent in the *H. robusta.* There are far more gene copies for the active molecules in *W. pigra* than in *H. robusta* (57 vs 24). *W. pigra* exceeds *H. robusta* in both kinds and gene number of active molecules. The gene copy of bioactive ingredients of *W. pigra also* exceeds that of *H. medicinalis*.

Table 2. The exploration of five class of active substances in *W. pigra, H. medicinalis* and *H. robusta*

Modes of action	Bioactive molecules	Gene copy (*H. robusta*)	Gene copy (*W. pigra*)	Gene copy (*H. medicinalis*)
Analgesic and anti-inflammatory effect	Antistasin	6	9	5
	Hirustasin	0	18	7
	Ghilanten	9	10	7
Extracellular matrix degradation	Hyaluronidase	1	4	5
Inhibition of platelet function	Saratin	1	4	1
Anticoagulant effect	hirudin	0	2	3
	Factor Xa inhibitor	2	1	1
	Therostasin	5	4	2
	Destabilase I [a]	0	5	2
Antimicrobial effect	Destabilase I [a]	0	5	2
Total		24	57	33

[a] Destabilase I, involved in anticoagulant effect and antimicrobial effect.

Gene expression analysis of *W. pigra*

We make full use of the available RNA-seq data to analyze the gene expression in *W. pigra*. We divided all genes of W. pigra into four parts (No expression, Low, Medium, High). The expression of these genes is shown in **Figure 5C**. These genes are involved in different molecular functions (**Figure 5D**). No expressed genes are enriched in the functions of ATPase activity, oxidoreductase activity, DNA−binding transcription factor activity, transposase activity and so on. Low expressed genes are enriched in the functions of ion transport, G protein−coupled receptor activity, carbohydrate binding, microtubule motor activity and so on. Medium expressed genes are enriched in the functions of nucleus, zinc ion binding, Rho guanyl−nucleotide exchange factor activity, protein dephosphorylation and so on. High expressed genes are enriched in the functions of translation, ribosome, serine−type endopeptidase inhibitor activity, enzyme inhibitor activity and so on. Finally, we examined the gene expression of bioactive peptide (**Figure 5E**). All kinds of bioactive peptides were expressed. Of these peptides, antistasin, therostasin, and hirudin have higher expression, while factor Xa inhibitor and ghilanten have lower expression. The result implies that these bioactive peptides may play different roles in the survival of *W. pigra*.

Discussion

"Medicinal leech" represents the leeches in the family hirudinidae of the order hirudinida. Medicinal leeches have been widely utilized in medical procedures for thousands of years and were approved by the US Food and Drug Administration in June, 2004 as a medical device due to their mechanically relieving venous congestion and delivering anti-coagulants[20, 21]. *W. pigra* is the most commonly available from Chinese commercial leech market. Although its importance in medicine and the significance of medicinal leeches in biological research, there is no genome data available for any species in the family hirudinidae until 2020. The genome of *H. robusta* has been sequenced to study bilaterian evolution in 2013. *H. robusta* is a leech of the family *glossiphoniidae*, which is very far from the family *hirudinidae* [22]. The genome of *H. medicinalis* in the family *hirudinidae* was just published during the review period of this study[8, 9]. In this study, we reported the genome of *W. pigra,* another medicinal leech in the family hirudinidae. We characterized the genome by analysis of gene synteny, gene family and the genome distribution of bioactive molecules and comparing it with *H. robusta* and *H. medicinalis*.

The results of the expansion and contraction of gene family revealed very clear different patterns among *W. pigra*, *H. medicinalis* and *H. robusta*. This suggests that the three leeches used different survival strategies to adapt to living environment. These results also suggest that although *W. pigra* and *H. medicinalis* both are medical leeches, they displayed different patterns of expanded and contracted families. Therefore, the features of one leech cannot simply be applied to another leech.

In the respect of active substances, we found a huge difference between *W. pigra* and *H. robusta* after systematic comparison of five classes of active substances. Hirudin, hirustasin, and destabilase I genes are absent in *H. robusta*. In contrast, all 9 common active molecules were found in *W. pigra*. There are two hirudin genes in *W. pigra*. Furthermore, two hirudin genes display different cysteine patterns in the protein sequence. The gene g17108 (hirudin_2) is a new kind of hirudin, which has not been reported before. The alignment of all available hirudin sequence shows a diversity of hirudin, which provides insight into the development of a new hirudin with more potent activity. Significantly, although W. pigra and H. medicinalis are both medicinal leeches, the gene copy of bioactive ingredients of *W. pigra* far exceeds that of *H. medicinalis*.

Conclusions

In summary, the genome of another medicinal leech (*W. pigra*) was reported in this study. The genomes of three leeches, *W. pigra, H. medicinalis* and *H. robusta,* show many differences in the respects of orthologous genes, gene synteny and gene family. Furthermore, *W. pigra* exceeds *H. robusta* in both kind and gene number of active substances, such as hirudin, hirustasin, and destabilase I genes. This study pointed out the differences in the genome of two medicinal leeches, *W. pigra* and *H. medicinalis* and provided insight into the exploration and development of bioactive molecules of medicinal leeches.

Methods

Sample preparation and genome and RNA-seq sequencing

Samples of *W. pigra* were collected from East Lake in Wuhan, the provincial capital Hubei, China. Animal care and handling were conducted in accordance with the stipulations of Ethics Committee of Kunming University. Genomic DNA was extracted from the whole body of *W. pigra* after cleaned off their gastric tracts and the blood. Two short paired-end (300 and 500 bp) and three mate-end (5, 8, and 10 Kbp, respectively) sequencing libraries were constructed with the standard protocol provided by Illumina (San Diego, United States), and then sequenced on an Illumina HiSeq™ 2000 platform. Low-quality and duplicated reads were filtered out through fastp (v0.20.0) software[23].

For transcriptome-based gene prediction, RNA was extracted from the whole body of *W. pigra* after cleaned off their gastric tracts and the blood. Prepared tissues (approximately 200 mg) were preserved in liquid nitrogen for RNA extraction. Total RNAs were purified with RNA Easy Kit (QIANGEN, German) according to the manufacturers' instructions. RNA yields and the quality were measured by agarose gel electrophoresis and spectrophotometer (Thermo, USA). Equal amounts of total RNA (20 μg, 50.8 μg/ml) purified from each tissue were separately stored and mRNA was isolated with Oligo-dT Purist Kit (TaKaRa, Japan) according to the standard protocol. All the libraries were prepared using the Illumina TruSeq RNA sample preparation kit (San Diego, United States) then were sequenced by Illumina HiSeq™ 2000 platform at Biomarker Technologies Co. Ltd. of Beijing, China.

Estimation of genome sizes and genome assembly

Genome sizes were estimated using JELLYFISH [24] and GenomeScope [25] with an optimal k-mer size (K-mer=23). Genome sizes were calculated from the following equation: Genome size = 23-mer_number / 23-mer_depth, where 23-mer_number is the total number of each unique 23-mer and 23-mer depth is the highest frequency that occurred. Consequently, the estimated genome size of *W. pigra* was ~ 162Mbp. By taking the estimated genome size as a reference, total sequence data accounted for ~100-fold coverage. The clean reads were used for de novo assembly by Platanus (v1.2.4)[26] with default parameters. Subsequently, intra-scaffold gaps were filled using the reads of short-insert libraries by gap_close command. The final assembled genome size was ~ 177 Mbp. The summary for assembly results are list in **Table 1**. Only scaffolds with lengths longer than 500 bp were used in further analyses.

Genome annotation

Homolog and de novo strategies were both applied to identify the repetitive sequence in the *W. pigra* genome. Software LTRfinder (v1.07)[27] and RepeatModeler (v1.0.11, http://www.repeatmasker.org/RepeatModeler) were used for ab initio prediction. The results obtained from these tools were combined to form a new repetitive sequence database. This database was then merged with Repbase [28, 29]. Repetitive sequences in the *W. pigra* genome were identified by homolog searching with the final merged database by RepeatMasker (v1.332)[30]. We identified 40 Mbp repetitive sequences, which accounted for 23% of the *W. pigra* assembled genome (**Table 1**). Protein coding genes were predicted using GeneMark-ES (v4.3.8) and AUGUSTUS (v3.3.0) implemented in the BRAKER2 pipeline[31, 32] using RNA-seq alignments as evidence. The RNA-seq bam files generated by HISAT2 [33,

34] were combined and fed into BRAKER. A total of 26,743 protein-coding genes were generated for the *W. pigra* genome.

All protein sequences from the BRAKER2 results were aligned to TrEMBL and UniProt [35] databases using BlastP at E-value $\leq 1e^{-5}$. Gene functions were also annotated using the InterProScan software[36-38] by searching publically available databases including Pfam[39, 40], PRINTS[41], ProDom[42] and SMART[43]. In summary, approximately 95% (25,496/26,743) of the genes were supported by at least one related function assignments from the public databases (TrEMBL, UniProt, and InterPro).

Comparative genomic analysis

To define gene families that descended from a single gene in the last common ancestor, we downloaded the protein-coding genes of *H. robusta, Lottia gigantea, Capitella teleta, Schmidtea mediterranea, Schistosoma mansoni* from NCBI. The protein-coding genes of *H. medicinalis* were downloaded from http://download.ripcm.com/hirudo_genome. The protein-coding gene of *W. pigra* were derived from BRAKER2. All proteins of the seven species were processed with OrthoFinder-Diamond (v1.1.10) to provide information about orthologous gene families. OrthoFinder is robust to incomplete models, differing gene lengths, and larger phylogenetic distances [44]. Gene families (orthogroups) in OrthoFinder are defined as homologous genes descended from a single gene from the last common ancestor of the species examined. It is assumed that a parental gene of each orthogroup was present in the common ancestor of the seven species investigated. We applied the likelihood model implemented in the software package CAFE (v4.1)[45] to identify the expanded and contracted gene family along each branch of the phylogenetic tree. The phylogenetic tree was constructed in the process of defining gene families.

Phylogenetic analysis of gene family

Protein sequences in the gene family were aligned using Clustal W [46] with fine adjustment by hand. Then the aligned sequences were used for phylogenetic analysis using MEGA X [47]. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [48]. The percentage of trees in which the associated taxa clustered together is shown next to the branches. Initial tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using a JTT model, and then selecting the topology with superior log likelihood value. The tree is drawn to scale, with branch lengths measured in the number of substitutions per site. The default parameters were used for sequence alignment, phylogenetic analysis.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The whole genome sequence data reported in this paper has been deposited in the Genome Warehouse in National Genomics Data Center [49], Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number GWHABJR00000000 that is publicly accessible at https://bigd.big.ac.cn/gwh.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by grants from the National Natural Science Foundation of China (31360516 and 31401142), the Joint Special Project of Universities in Yunnan (2017FH001-004), Yunnan Provincial Training Programs of Youth Leader in Academic and Technical Reserve Talent (2018HB101), Yunnan Provincial Ten Thousand People Plan, the start-up fund of Kunming university of science and technology (KKZ3201927005), and Yunnan Fundamental Research Projects (2019FB050).

Authors' contributions

ZCL and SXD conceived the study and designed experiments. LT, SXD and DJK performed the experiments. SXD, XT and ZCL analyzed the data. XRT, XXB, YS, and YQZ contributed reagents/materials/analysis tools. ZCL, LT and SXD wrote and revised the paper.

Acknowledgements

We thank our colleagues, Drs. Gong-Hua Li and Wen-Xing Li for helpful comments on the manuscript.

References

1. Joslin J, Biondich A, Walker K, Zanghi N: **A Comprehensive Review of Hirudiniasis: From Historic Uses of Leeches to Modern Treatments of Their Bites.** *Wilderness Environ Med* 2017, **28:**355-361.
2. Zaidi SM, Jameel SS, Zaman F, Jilani S, Sultana A, Khan SA: **A systematic overview of the medicinal importance of sanguivorous leeches.** *Altern Med Rev* 2011, **16:**59-65.
3. Zhang Y: **Why do we study animal toxins?** *Zoological research* 2015, **36:**183.
4. Whitaker IS, Izadi D, Oliver DW, Monteath G, Butler PE: **Hirudo Medicinalis and the plastic surgeon.** *Br J Plast Surg* 2004, **57:**348-353.
5. Kuo DH, Lai YT: **On the origin of leeches by evolution of development.** *Dev Growth Differ* 2019, **61:**43-57.

6. Liu Z, Tong X, Su Y, Wang D, Du X, Zhao F, Wang D, Zhao F: **In-depth profiles of bioactive large molecules in saliva secretions of leeches determined by combining salivary gland proteome and transcriptome data.** *J Proteomics* 2019, **200:**153-160.
7. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH, Larsson T, Lv J, Arendt D, et al: **Insights into bilaterian evolution from three spiralian genomes.** *Nature* 2013, **493:**526-531.
8. Babenko VV, Podgorny OV, Manuvera VA, Kasianov AS, Manolov AI, Grafskaia EN, Shirokov DA, Kurdyumov AS, Vinogradov DV, Nikitina AS, et al: **Draft genome sequences of Hirudo medicinalis and salivary transcriptome of three closely related medicinal leeches.** *BMC Genomics* 2020, **21:**331.
9. Kvist S, Manzano-Marin A, de Carle D, Trontelj P, Siddall ME: **Draft genome of the European medicinal leech Hirudo medicinalis (Annelida, Clitellata, Hirudiniformes) with emphasis on anticoagulants.** *Sci Rep* 2020, **10:**9885.
10. Phillips AJ, Siddall ME: **Poly-paraphyly of Hirudinidae: many lineages of medicinal leeches.** *BMC Evol Biol* 2009, **9:**246.
11. Khan MS, Guan DL, Kvist S, Ma LB, Xie JY, Xu SQ: **Transcriptomics and differential gene expression in Whitmania pigra (Annelida: Clitellata: Hirudinida: Hirudinidae): Contrasting feeding and fasting modes.** *Ecol Evol* 2019, **9:**4706-4719.
12. Liu Z, Wang Y, Tong X, Su Y, Yang L, Wang D, Zhao Y: **De novo assembly and comparative transcriptome characterization of Poecilobdella javanica provide insight into blood feeding of medicinal leeches.** *Mol Omics* 2018, **14:**352-361.
13. Liu Z, Zhao F, Tong X, Liu K, Wang B, Yang L, Ning T, Wang Y, Zhao F, Wang D, Wang D: **Comparative transcriptomic analysis reveals the mechanism of leech environmental adaptation.** *Gene* 2018, **664:**70-77.
14. Dong H, Ren JX, Wang JJ, Ding LS, Zhao JJ, Liu SY, Gao HM: **Chinese Medicinal Leech: Ethnopharmacology, Phytochemistry, and Pharmacological Activities.** *Evid Based Complement Alternat Med* 2016, **2016:**7895935.
15. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al: **MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity.** *Nucleic Acids Res* 2012, **40:**e49.
16. Muller C, Haase M, Lemke S, Hildebrandt JP: **Hirudins and hirudin-like factors in Hirudinidae: implications for function and phylogenetic relationships.** *Parasitol Res* 2017, **116:**313-325.
17. Markwardt F: **Hirudin as alternative anticoagulant–a historical review.** *Semin Thromb Hemost* 2002, **28:**405-414.
18. Hildebrandt JP, Lemke S: **Small bite, large impact-saliva and salivary molecules in the medicinal leech, Hirudo medicinalis.** *Naturwissenschaften* 2011, **98:**995-1008.
19. Hibsh D, Schori H, Efroni S, Shefi O: **De novo transcriptome assembly databases for the central nervous system of the medicinal leech.** *Sci Data* 2015, **2:**150015.

20. Derganc M, Zdravic F: **Venous congestion of flaps treated by application of leeches.** *Br J Plast Surg* 1960, **13:**187-192.
21. Rados C: **Beyond bloodletting: FDA gives leeches a medical makeover.** *FDA Consum* 2004, **38:**9.
22. Apakupakul K, Siddall ME, Burreson EM: **Higher level relationships of leeches (Annelida: Clitellata: Euhirudinea) based on morphology and gene sequences.** *Mol Phylogenet Evol* 1999, **12:**350-359.
23. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ preprocessor.** *Bioinformatics* 2018, **34:**i884-i890.
24. Marcais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of occurrences of k-mers.** *Bioinformatics* 2011, **27:**764-770.
25. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC: **GenomeScope: fast reference-free genome profiling from short reads.** *Bioinformatics* 2017, **33:**2202-2204.
26. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H, et al: **Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.** *Genome Res* 2014, **24:**1384-1395.
27. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons.** *Nucleic Acids Res* 2007, **35:**W265-268.
28. Bao W, Kojima KK, Kohany O: **Repbase Update, a database of repetitive elements in eukaryotic genomes.** *Mob DNA* 2015, **6:**11.
29. Jurka J: **Repeats in genomic DNA: mining and meaning.** *Curr Opin Struct Biol* 1998, **8:**333-337.
30. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in genomic sequences.** *Curr Protoc Bioinformatics* 2009, **Chapter 4:**Unit 4 10.
31. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M: **Whole-Genome Annotation with BRAKER.** *Methods Mol Biol* 2019, **1962:**65-95.
32. Lomsadze A, Burns PD, Borodovsky M: **Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm.** *Nucleic Acids Res* 2014, **42:**e119.
33. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory requirements.** *Nat Methods* 2015, **12:**357-360.
34. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL: **Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype.** *Nat Biotechnol* 2019, **37:**907-915.
35. UniProt Consortium T: **UniProt: the universal protein knowledgebase.** *Nucleic Acids Res* 2018, **46:**2699.
36. Zdobnov EM, Apweiler R: **InterProScan–an integration platform for the signature-recognition methods in InterPro.** *Bioinformatics* 2001, **17:**847-848.
37. Mulder N, Apweiler R: **InterPro and InterProScan: tools for protein sequence classification and comparison.** *Methods Mol Biol* 2007, **396:**59-70.

38. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, et al: **InterProScan 5: genome-scale protein function classification.** *Bioinformatics* 2014, **30:**1236-1240.
39. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, et al: **The Pfam protein families database in 2019.** *Nucleic Acids Res* 2019, **47:**D427-D432.
40. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al: **The Pfam protein families database: towards a more sustainable future.** *Nucleic Acids Res* 2016, **44:**D279-285.
41. Attwood TK, Coletta A, Muirhead G, Pavlopoulou A, Philippou PB, Popov I, Roma-Mateo C, Theodosiou A, Mitchell AL: **The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012.** *Database (Oxford)* 2012, **2012:**bas019.
42. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains.** *Brief Bioinform* 2002, **3:**246-251.
43. Letunic I, Bork P: **20 years of the SMART protein domain annotation resource.** *Nucleic Acids Res* 2018, **46:**D493-D496.
44. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy.** *Genome Biol* 2015, **16:**157.
45. De Bie T, Cristianini N, Demuth JP, Hahn MW: **CAFE: a computational tool for the study of gene family evolution.** *Bioinformatics* 2006, **22:**1269-1271.
46. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, et al: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23:**2947-2948.
47. 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.
48. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8:**275-282.
49. Members BIGDC: **Database Resources of the BIG Data Center in 2019.** *Nucleic Acids Res* 2019, **47:**D8-D14.

Figures

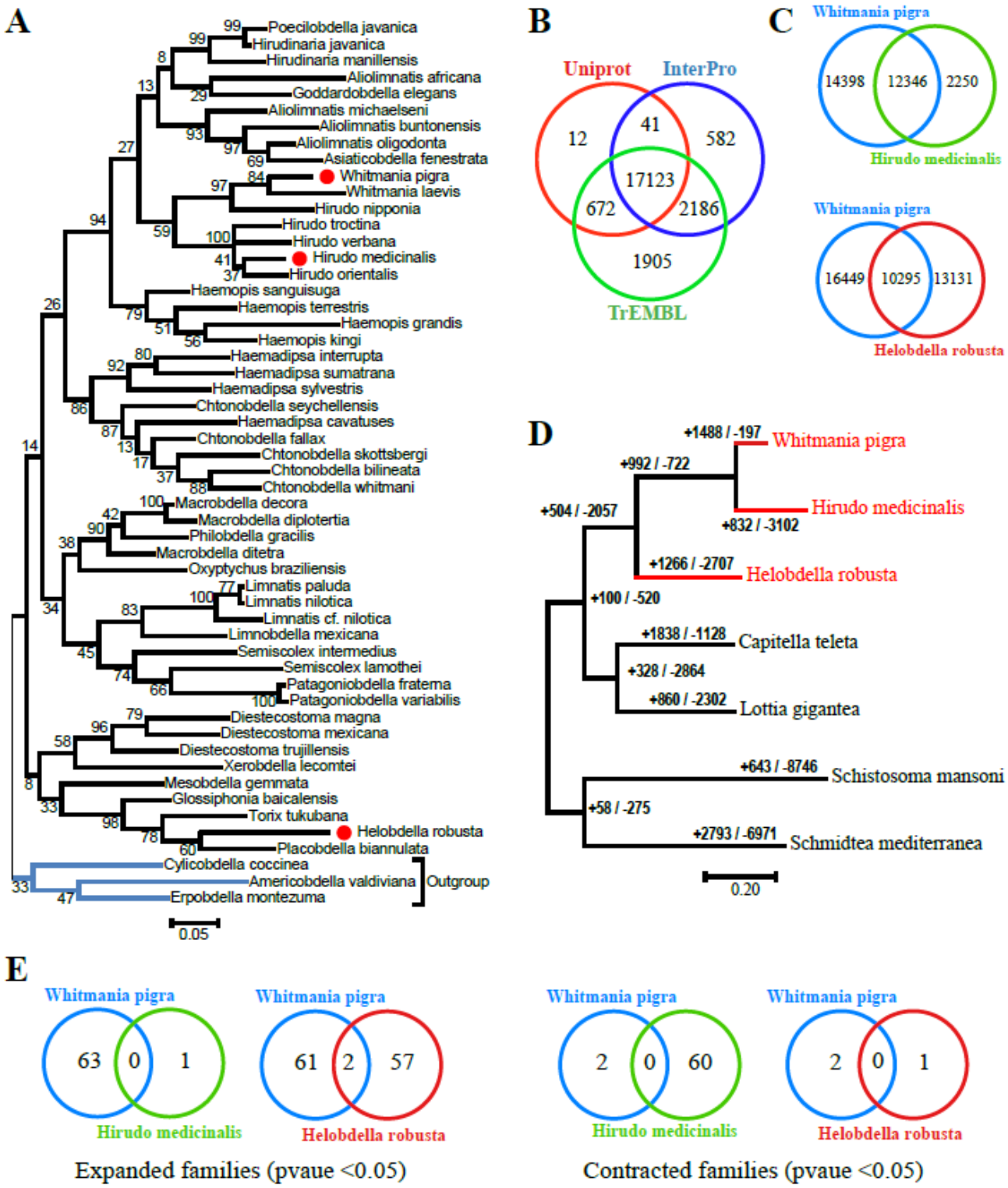

Figure 1

Genome annotation and evolution of W. pigra compared with H. robusta and H. medicinalis. A) Phylogenetic analysis of leech species by Maximum Likelihood method based on COI genes. Highlighted with red dots correspond to three leeches compared with in our study. Posterior probabilities are assigned to the node; B) the predicted protein-coding genes with matching entries in the three popular public databases; C) the Venn diagram showed the orthologous genes between W. pigra and H. medicinalis (top

panel), and between W. pigra and H. robusta (bottom panel); D) Gene expansion and contraction in the W. pigra genome. The number of expanded (+) and contracted (-) gene families are shown along branches and nodes. E) the Venn diagram showed the number of expanded and contracted gene families between W. pigra and H. medicinalis, and between W. pigra and H. robusta.

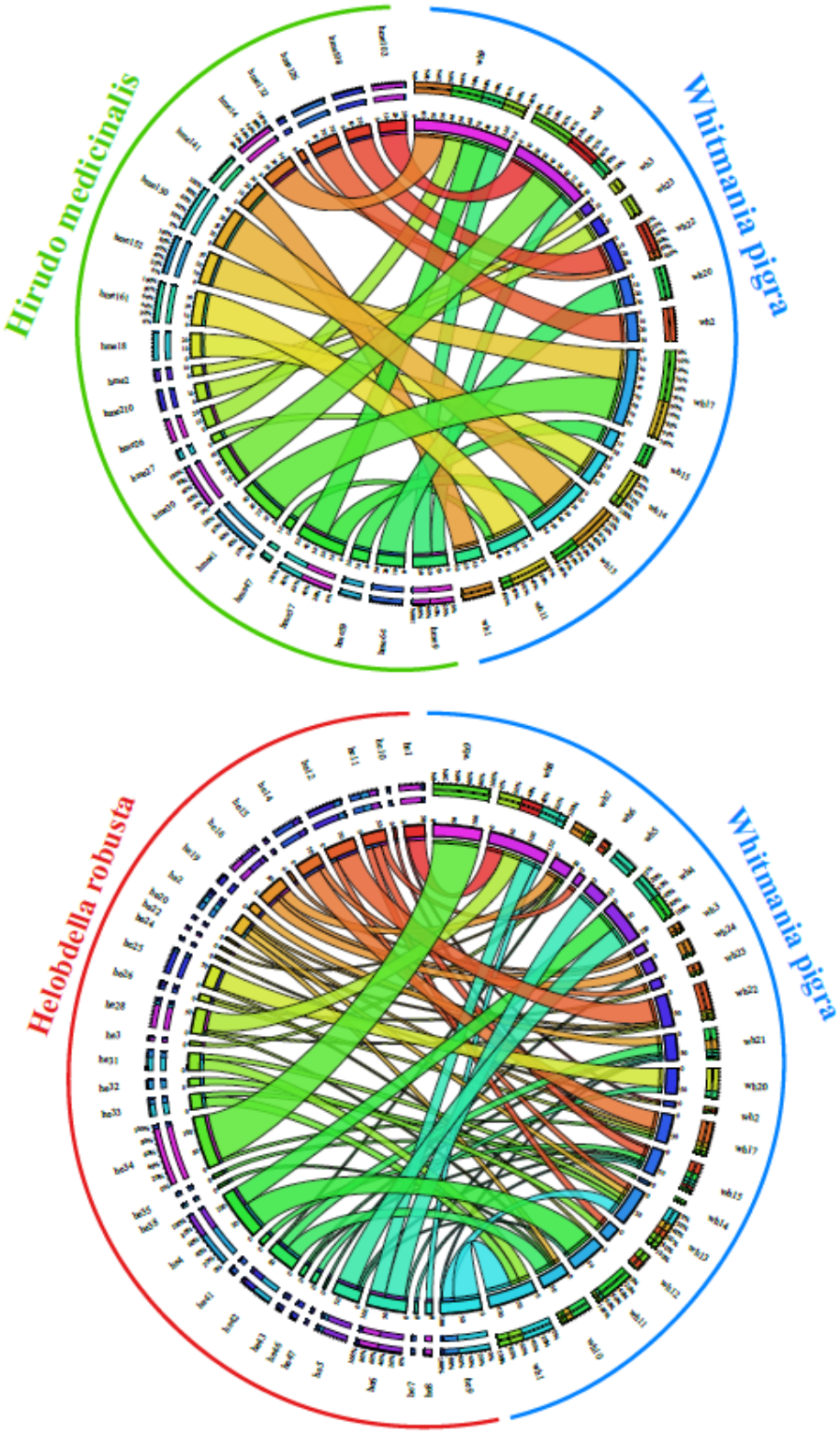

Figure 2

Syntenic relationships between W. pigra and H. medicinalis, and between W. pigra and H. robusta. The top panel represents the syntenic relationships between W. pigra and H. medicinalis. The bottom panel shows the syntenic relationships between W. pigra and H. robusta. The scaffolds will be connected if they share similar genes. The width of link represents the number of shared genes.

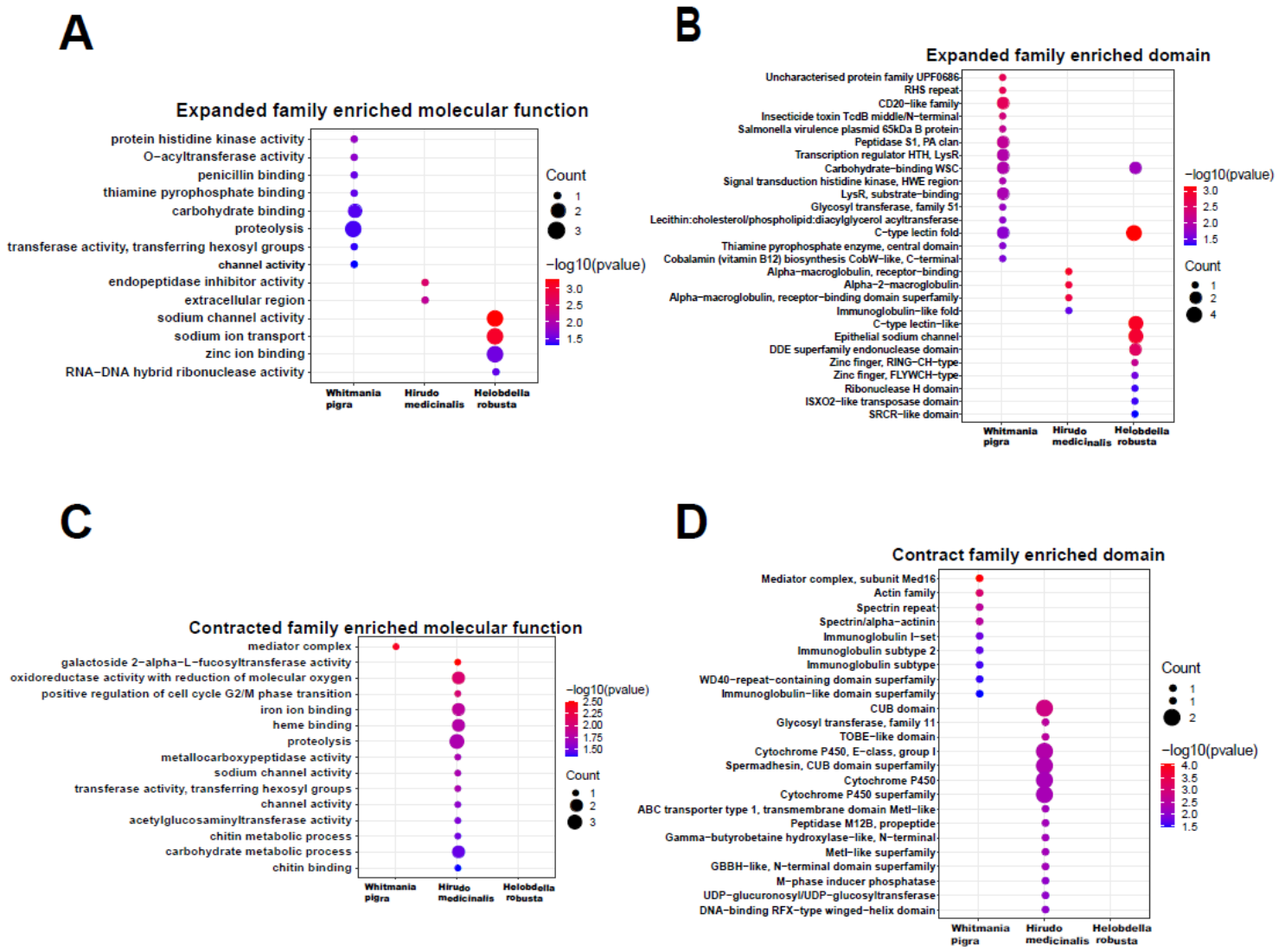

Figure 3

Enrichment analysis of expanded and contracted gene families between the three species using GO terms and interPro domains. GO terms (A) and interPro domains (B) were enriched by expanded gene families in W. pigra, H. medicinalis and H. robusta; GO terms (C) and interPro domains (D) were enriched by contracted gene families in W. pigra, H. medicinalis and H. robusta.

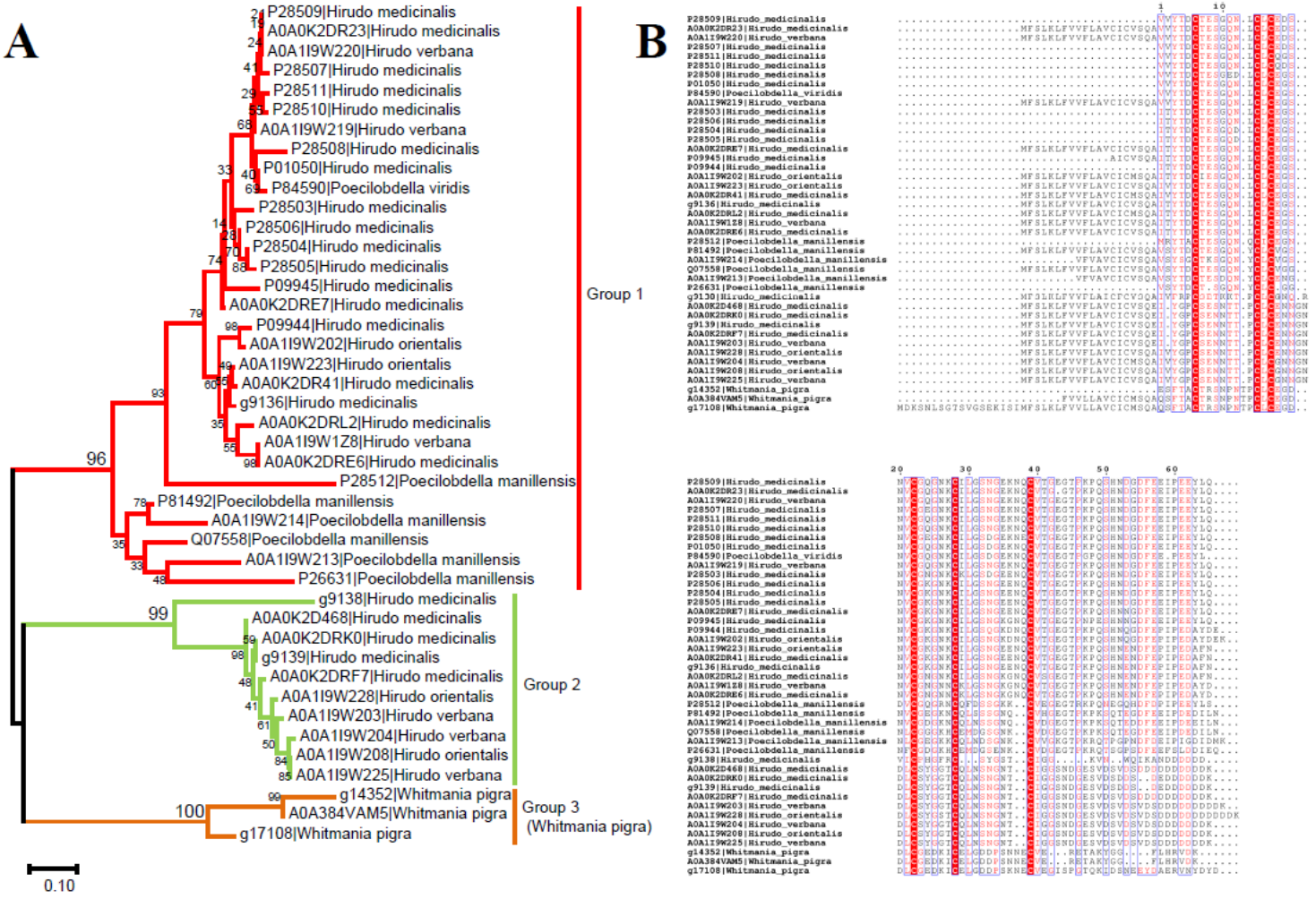

Figure 4

Sequence analysis of the hirudin gene family. A) Phylogenetic analysis of hirudin gene family from the species of family hirudinidae. The tree was inferred by using the Maximum Likelihood method and JTT matrix-based model Likelihood method. B) Multiple alignments of the amino acid sequences of hirudin proteins.

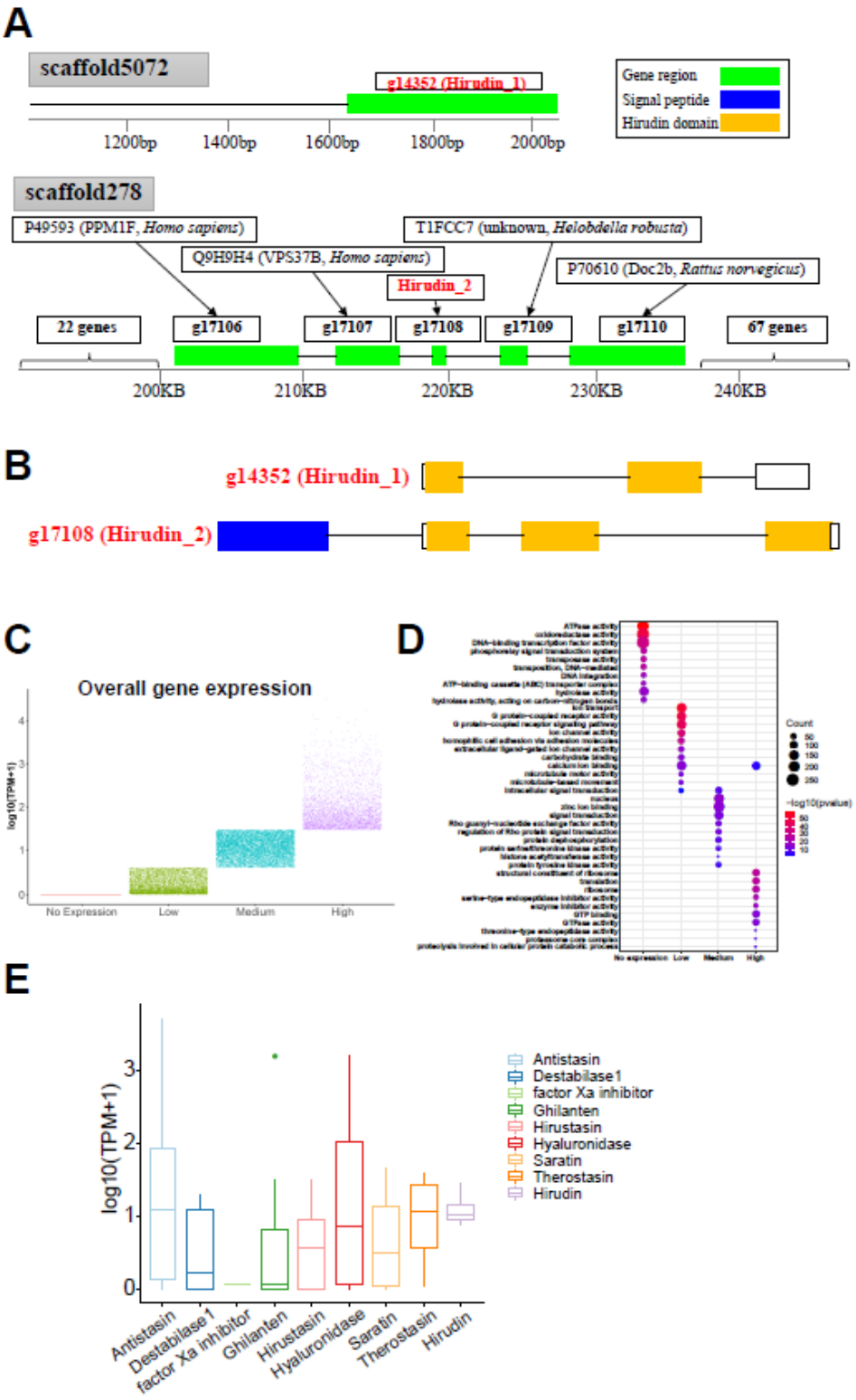

Figure 5

Detailed analysis of hirudin genes and gene expression in W. pigra. A) Genome-wide distribution of hirudin genes; B) Gene structure of hirudin genes; C) Jitter plot shown overall gene expression in W. pigra. D) Enriched molecular functions by the four parts of genes, respectively. E) Gene expression of different classes of bioactive peptides.

Supplementary Files

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

- Supplementaryinformation.pdf