of Bioactive Ingredients

Lei Tong^{1,#}, Shao-Xing Dai^{2,#}, De-Jun Kong^{1,#}, Peng-Peng Yang², Xin Tong²,
Xiang-Rong Tong¹, Xiao-Xu Bi¹, Yuan Su¹, Yu-Qi Zhao³, Zi-Chao Liu^{1,*}

¹Engineering Research Center for Exploitation and Utilization of Leech Resources in
Universities of Yunnan Province, School of Agronomy and Life Sciences, Kunming
University, Kunming 650214, China

²Yunnan Key Laboratory of Primate Biomedicine Research, Institute of Primate
Translational Medicine, Kunming University of Science and Technology, Kunming
650500, China

³Department of Integrative Biology and Physiology, University of California, Los
Angeles, Los Angeles, CA 90095, USA

Authorsemail:

Lei Tong: 1501159559@qq.com

Shao-Xing Dai: daishx08@gmail.com

De-Jun Kong: kongdj05@hotmail.com

Peng-Peng Yang: yangpp@lpbr.cn

Xin Tong: tongxin0013@163.com

Xiang-Rong Tong: txiangrong@163.com

Xiao-Xu Bi: 279781667@qq.com

Yuan Su: suyuankm@163.com

Yu-Qi Zhao: zhaoyuqi616@gmail.com

Zi-Chao Liu: abclzc@aliyun.com

#These authors contributed equally to this work

*Corresponding author: Zi-Chao Liu

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. The discovery of bioactive ingredients is hindered heavily due to the lack of the genome of medicinal leeches.

Results

To facilitate this discovery and the research of medicinal leeches, we reported the first genome of the medicinal leech (*Whitmania pigra*). 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. There are only 9799 (57%) orthologous genes between *W. pigra* and *Helobdella robusta*. And only 16% genes in *W. pigra* showed syntenic arrangement with *H. robusta* revealed by gene synteny analysis. Furthermore, *W. pigra* and *H. robusta* expanded different gene families enriched in the ‘calcium ion’ and ‘sodium ion’ related functions, respectively. By inspecting genome distribution and gene structure of hirudin, we identified a new hirudin gene g17108 (hirudin_2) with different cysteine pattern. At last, we systematically explored and compared the active substances in the genomes of two leeches. The results showed that *W. pigra* exceeds *H. robusta* in both kinds and gene number of active molecules.

Conclusions

This study reported the first genome of the medicinal leech (*W. pigra*), which provides an important genome resource and new insight into the exploration and development of bioactive molecules of medicinal leeches.

Keywords: *Whitmania pigra*, Genome, Bioactive Ingredients, *Helobdella robusta*

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]. 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 a marine species, 90 terrestrial and the remainder freshwater taxa.

Although a huge diversity and closely relationship to people, we know little about the genome of leeches. Up to now, only one species *H. robusta* has been sequenced for study bilaterian evolution in 2013 [7]. *H. robusta* is a leech of the family glossiphoniidae. Its early development has been studied extensively. For another important family hirudinidae, the genome data were totally absent. 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.

The lack of genomic data has greatly limited on the exploration of bioactive ingredients. In this study, we reported the first genome of the medicinal leech *W. pigra*. *W. pigra*, an Asian freshwater leech, is non-blood feeding, despite the placement of this genus within the family hirudinidae [8] which 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[9-11]. We conducted an analysis of gene synteny between the two leeches *H. robusta* and *W. pigra*. Then we analyzed the expansion and contraction of gene family among six related species (*H. robusta*, *Lottia gigantea*, *Capitella teleta*, *Schmidtea mediterranea*, *Schistosoma mansoni*, *W. pigra*). 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 *W. pigra* and *H. robusta*. This study pointed out the differences in the genome of two leeches *W. pigra* 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 1A**). We identified 9799 orthologous genes between *W. pigra* and *H. robusta* using reciprocal best blast hits (RBHs) method (**Figure 1B**). There are a large proportion of genes (7259) not assigned as orthologous genes.

Table 1. Summary for genome sequencing, assembly and annotation

	<i>H. robusta</i>	<i>W. pigra</i>
Size of genome assembly	228 Mbp	177 Mbp
Num. of Scaffolds	1,993	10,050
Num. of Scaffolds (> 2Kbp)	1124	3495
Scaffold N50	3,060 Kbp	728 Kbp
Total reads	3,176,156	118,388,619
Reads mapping to genome (%)	2,839,951 (89%)	112,480,685 (95%)
Sequencing read coverage depth	7.92X	100X
Repetitive content (%)	33	23
GC (%)	33	35
Num. of predicted genes	23,400	26,743
Protein length (aa)	376	438

Mean exon length	203 bp	205bp
Mean intron length	526 bp	391bp
Mean number of exons per gene	6.1	6.4

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 between the genomes of *W. pigra* and *H. robusta*

Above result showed that *W. pigra* only has 57.4% (9799/17058) orthologous genes in *H. robusta*. To further compare the genome similarity between the two leeches, we performed a careful analysis of syntenic blocks between *W. pigra* and *H. robusta* using MCScanX[12]. A total of 493 syntenic blocks, including 4416 gene pairs, were identified (**Figure 2**). Each syntenic block shares 15 genes in average. Scaffold he490 (*H. robusta*) and wh1108 (*W. pigra*) have largest syntenic regions, including 92 similar genes. Only 16% (4416/26743) genes in *W. pigra* show syntenic arrangement with *H. robusta*. It suggests that there are some big differences between genomes of *W. pigra* and *H. robusta*.

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 six species: *H. robusta*, *Lottia gigantea*, *Capitella teleta*, *Schmidtea mediterranea*, *Schistosoma mansoni*, *W. pigra*. We compared the predicted proteomes of six species, yielding a total of 12367 orthologous gene families that comprised 91225 genes. We found 1511 and 965 gene families expanded in *W. pigra* and *H. robusta*, respectively (**Figure 1C**). Of these families, there are 54 and 68 are

rapidly evolving families ($P < 0.05$) in *W. pigra* and *H. robusta*, respectively (**Figure 1D**). These rapidly evolving families are species-specific and not overlap between the two species (**Figure 1D**). To reveal the molecular function and pathway of these rapidly evolving families, we performed enrichment analyses by gene ontology terms, KEGG pathways and InterPro domains. We found some enriched functions such as calcium ion binding (GO:0005509, $P = 1e-52$), homophilic cell adhesion (GO:0007156, $P = 3e-51$), carbohydrate binding (GO:0030246, $P = 1e-13$), oxidoreductase activity (GO:0016627, $P = 5e-13$) and metalloendopeptidase activity (GO:0004222, $P = 3e-12$) in *W. pigra* (**Figure 3A**). Corresponding to these functions, many related domains were enriched by the expansive genes, for example, “Cadherin-like superfamily”, “Cadherin-like”, “Cadherin, N-terminal”, “Carbohydrate-binding WSC”, “Peptidase M12B, ADAM/reprolysin” and so on (**Figure 3B**). In contrast, rather than calcium ion binding function, multiple sodium related functions were found to be enriched in *H. robusta*, such as “sodium channel activity”, and “sodium ion transport” (**Figure 3C and D**). Calcium functions as an effector signaling molecule by modulating the function of intracellular enzymes and voltage-gated ion channels. Sodium channels are the main trigger for electrical signals in all animals, which are the basis of neural activity and muscle contraction. In addition, the enriched KEGG pathways are also different between the two species (**Figure S2**). It implies the two species may take different adaptive strategies. And the different functions and domains and pathways are potentially related to environmental adaptation and phytochemical properties of the two 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[13]. Hirudin may be useful in the therapy of thrombosis because of its specific antithrombin effects[14]. 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**). We identified 39 hirudin-like sequences from protein database UniProt through BLASTP. Then all 41 hirudin sequence, including g14352 and g17108, 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 >90. 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 cysteine. 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 hirudingenes. Furthermore, gene structures of the two hirudingenes 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**). The result of transcriptome showed that the two hirudingenes had very higher expression level than average (**Figure 5C**). It implies the two hirudingenes play an important role in survival of *W. pigra*.

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

There are more than 20 bioactive substances were identified from leeches, such as Antistasin, hirustasin, ghilantens, hirudin[2, 6, 15, 16]. These molecules have analgesic, anti-inflammatory, anticoagulant, platelet inhibitory, thrombin regulatory functions, and so on. *W. pigra* and *H. robusta* are belong 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* 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* (45 vs 23). *W.*

pigra exceeds *H. robusta* in both kinds and gene number of active molecules.

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

Modes of action	Bioactive molecules	Gene copy	Gene copy
Analgesic and anti-inflammatory effect	Antistasin	5	5
	Hirustasin	0	10
	Ghilanten	9	7
Extracellular matrix degradation	Hyaluronidase	1	4
Inhibition of platelet	Saratin	1	4
Anticoagulant effect	hirudin	0	2
	Factor Xa inhibitor	2	1
	Therostasin	5	2
	Destabilase I ^a	0	5
Antimicrobial effect	Destabilase I ^a	0	5
Total		23	45

^a Destabilase I, involved in anticoagulant effect and antimicrobial effect.

Discussion

"Medicinal leech" represents the blood-feeding leech 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[17, 18]. Although its importance in medicine and the

significance of medicinal leeches in biological research, there are no genome data available for any species in the family hirudinidae. So far, only the genome of *H. robusta* has been sequenced for study bilaterian evolution in 2013. *H. robusta* is a leech of the family *glossiphoniidae*, which is very far from the family *hirudinidae*[19].The absence of genomic data has seriously restricted the exploration of bioactive ingredients from the species in the family hirudinidae. In this study, we reported the genome of *W. pigra* in the family hirudinidae. We characterized the genome by analysis of gene synteny, gene family and the genome distribution of bioactive molecules and comparing with *H. robusta* and other related species.

The results of the expansion and contraction of gene family revealed different patterns between *W. pigra* and *H. robusta*. The expansive gene families of *W. pigra* are enriched in the ‘calcium ion’ related functions. In contrast, multiple ‘sodium ion’ related functions were found to be enriched for the expansive gene families in *H. robusta*. Accordingly, the enriched protein domains are also different between the two species. This suggests that the two species used different survival strategies to adapt living environment. These results also suggest that although they both are leeches, the results 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 provide insight into the development of a new hirudin with more potent activity.

Conclusions

In summary, the first genome of the medicinal leech (*W. pigra*) was reported in this study. The genomes of two leeches *W. pigra* 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 leeches *W. pigra* and *H. robusta* 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 [20].

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 (QIAGEN, 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 [21] and GenomeScope [22] with an optimal k-mer size (K-mer=23). Genome sizes were calculated from the following equation: $\text{Genome size} = \text{23-mer_number} / \text{23-mer_depth}$, 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 ~ 162 Mbp. 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)[23] 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 the **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)[24] and RepeatModeler (v1.0.11, <http://www.repeatmasker.org/RepeatModeler>) were used for ab initio prediction. The results obtained from these softwares were combined to form a new repetitive sequence database. This database was then merged with Repbase [25, 26]. Repetitive sequences in the *W. pigra* genome were identified by homolog searching with the final merged database by RepeatMasker (v1.332)[27]. 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[28, 29] using RNA-seq alignments as evidence. The RNA-seq bam files generated by HISAT2 [30, 31] 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 [32] databases using BlastP at E-value $\leq 1e^{-5}$. Gene functions were also annotated using the InterProScan software[33-35] by searching publically available databases including Pfam[36, 37], PRINTS[38], ProDom[39] and SMART[40]. 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 gene of *W. pigra* were derived from BRAKER2. All proteins of the six 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 [41]. 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 six species investigated. We applied the likelihood model implemented in the software package CAFE (v4.1)[42] 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 [43] with fine adjustment by hand. Then the aligned sequences were used for phylogenetic analysis using MEGA X [44]. The evolutionary history was inferred by using the Maximum Likelihood method and JTT matrix-based model [45]. 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 have been deposited in the Genome Warehouse in National Genomics Data Center[46], Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number GWHABJR000000000 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

Province (2017FH001-004), 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, DBW 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.

- 357 7. Simakov O, Marletaz F, Cho SJ, Edsinger-Gonzales E, Havlak P, Hellsten U, Kuo DH,
358 Larsson T, Lv J, Arendt D, et al: **Insights into bilaterian evolution from three spiralian**
359 **genomes.*Nature* 2013, **493**:526-531.**
- 360 8. Phillips AJ, Siddall ME: **Poly-paraphyly of Hirudinidae: many lineages of medicinal**
361 **leeches.*BMC Evol Biol* 2009, **9**:246.**
- 362 9. Khan MS, Guan DL, Kvist S, Ma LB, Xie JY, Xu SQ: **Transcriptomics and differential gene**
363 **expression in Whitmania pigra (Annelida: Clitellata: Hirudinida: Hirudinidae):**
364 **Contrasting feeding and fasting modes.*Ecol Evol* 2019, **9**:4706-4719.**
- 365 10. Liu Z, Wang Y, Tong X, Su Y, Yang L, Wang D, Zhao Y: **De novo assembly and comparative**
366 **transcriptome characterization of Poecilobdella javanica provide insight into blood**
367 **feeding of medicinal leeches.*Mol Omics* 2018, **14**:352-361.**
- 368 11. Liu Z, Zhao F, Tong X, Liu K, Wang B, Yang L, Ning T, Wang Y, Zhao F, Wang D, Wang D:
369 **Comparative transcriptomic analysis reveals the mechanism of leech environmental**
370 **adaptation.*Gene* 2018, **664**:70-77.**
- 371 12. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al:
372 **MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and**
373 **collinearity.*Nucleic Acids Res* 2012, **40**:e49.**
- 374 13. Muller C, Haase M, Lemke S, Hildebrandt JP: **Hirudins and hirudin-like factors in**
375 **Hirudinidae: implications for function and phylogenetic relationships.*Parasitol Res* 2017,
376 **116**:313-325.**
- 377 14. Markwardt F: **Hirudin as alternative anticoagulant--a historical review.*Semin Thromb*
378 *Hemost* 2002, **28**:405-414.**
- 379 15. Hildebrandt JP, Lemke S: **Small bite, large impact-saliva and salivary molecules in the**
380 **medicinal leech, Hirudo medicinalis.*Naturwissenschaften* 2011, **98**:995-1008.**
- 381 16. Hibsh D, Schori H, Efroni S, Shefi O: **De novo transcriptome assembly databases for the**
382 **central nervous system of the medicinal leech.*Sci Data* 2015, **2**:150015.**
- 383 17. Derganc M, Zdravic F: **Venous congestion of flaps treated by application of leeches.*Br J*
384 *Plast Surg* 1960, **13**:187-192.**
- 385 18. Rados C: **Beyond bloodletting: FDA gives leeches a medical makeover.*FDA Consum* 2004,
386 **38**:9.**

- 387 19. Apakupakul K, Siddall ME, Burrenson EM: **Higher level relationships of leeches (Annelida:**
388 **Clitellata: Euhirudinea) based on morphology and gene sequences.***Mol Phylogenet Evol*
389 1999, **12**:350-359.
- 390 20. Chen S, Zhou Y, Chen Y, Gu J: **fastp: an ultra-fast all-in-one FASTQ**
391 **preprocessor.***Bioinformatics* 2018, **34**:i884-i890.
- 392 21. Marçais G, Kingsford C: **A fast, lock-free approach for efficient parallel counting of**
393 **occurrences of k-mers.***Bioinformatics* 2011, **27**:764-770.
- 394 22. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, Schatz MC:
395 **GenomeScope: fast reference-free genome profiling from short reads.***Bioinformatics* 2017,
396 **33**:2202-2204.
- 397 23. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M,
398 Nagayasu E, Maruyama H, et al: **Efficient de novo assembly of highly heterozygous**
399 **genomes from whole-genome shotgun short reads.***Genome Res* 2014, **24**:1384-1395.
- 400 24. Xu Z, Wang H: **LTR_FINDER: an efficient tool for the prediction of full-length LTR**
401 **retrotransposons.***Nucleic Acids Res* 2007, **35**:W265-268.
- 402 25. Bao W, Kojima KK, Kohany O: **Repbase Update, a database of repetitive elements in**
403 **eukaryotic genomes.***Mob DNA* 2015, **6**:11.
- 404 26. Jurka J: **Repeats in genomic DNA: mining and meaning.***Curr Opin Struct Biol* 1998,
405 **8**:333-337.
- 406 27. Tarailo-Graovac M, Chen N: **Using RepeatMasker to identify repetitive elements in**
407 **genomic sequences.***Curr Protoc Bioinformatics* 2009, **Chapter 4**:Unit 4 10.
- 408 28. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M: **Whole-Genome Annotation with**
409 **BRAKER.***Methods Mol Biol* 2019, **1962**:65-95.
- 410 29. Lomsadze A, Burns PD, Borodovsky M: **Integration of mapped RNA-Seq reads into**
411 **automatic training of eukaryotic gene finding algorithm.***Nucleic Acids Res* 2014, **42**:e119.
- 412 30. Kim D, Langmead B, Salzberg SL: **HISAT: a fast spliced aligner with low memory**
413 **requirements.***Nat Methods* 2015, **12**:357-360.
- 414 31. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL: **Graph-based genome alignment and**
415 **genotyping with HISAT2 and HISAT-genotype.***Nat Biotechnol* 2019, **37**:907-915.
- 416 32. UniProt Consortium T: **UniProt: the universal protein knowledgebase.***Nucleic Acids Res*

2018, **46**:2699.

33. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro.***Bioinformatics* 2001, **17**:847-848.

34. Mulder N, Apweiler R: **InterPro and InterProScan: tools for protein sequence classification and comparison.***Methods Mol Biol* 2007, **396**:59-70.

35. 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.

36. 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.

37. Finn RD, Coghill 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.

38. 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.

39. 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.

40. Letunic I, Bork P: **20 years of the SMART protein domain annotation resource.***Nucleic Acids Res* 2018, **46**:D493-D496.

41. Emms DM, Kelly S: **OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy.***Genome Biol* 2015, **16**:157.

42. 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.

43. 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.

44. 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.

45. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8**:275-282.

Members BIGDC: **Database Resources of the BIG Data Center in 2019.** *Nucleic Acids Res* 2019, **47**:D8-D14.

Figure legends

Figure 1. Genome annotation and evolution of *W. pigra* compared with *H. robusta*. **A)** Phylogenetic analysis of leech species by Maximum Likelihood method based on COI genes. Highlighted with red dots correspond to two 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. robusta*; **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 the two species.

Figure 2. Syntenic relationships between *W. pigra* and *H. robusta*. The left arc represents *H. robusta* and the right arc shows the syntenic blocks of *W. pigra*. 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 gene families between the two species using GO terms and interPro domains. GO terms (**A**) and interPro domains (**B**) were

470 enriched in *W. pigra*; GO terms (C) and interPro domains (D) were enriched in *H. robusta*.

471 **Figure 4. Sequence analysis of the hirudin gene family.** A) Phylogenetic analysis of
472 hirudin gene family from the species of family hirudinidae. The tree was inferred by
473 using the Maximum Likelihood method and JTT matrix-based model Likelihood
474 method. B) Multiple alignments of the amino acid sequences of hirudin proteins.

475 **Figure 5. Detailed analysis of hirudin genes in *W. pigra*.** A) Genome-wide
476 distribution of hirudin genes; B) Gene structure of hirudin genes; C) Transcript levels
477 of hirudin genes.