A Draft Genome for *Hirudo verbana*, the Medicinal Leech

Authors and Affiliations

Riley T. Paulsen[1,2], Diing D.M. Agany[3], Jason Petersen[4,5], Christel M. Davis[4], Erik A. Ehli[4,6], Etienne Gnimpieba[2,3], and Brian D. Burrell[1,2*]

[1]Division of Basic Biomedical Sciences and Center for Brain and Behavior Research, Sanford School of Medicine, University of South Dakota, Vermillion, SD, USA

[2]Center for Brain and Behavior Research, University of South Dakota, SD, USA

[3]Biomedical Engineering Program, University of South Dakota, Sioux Falls, SD, USA

[4]Avera Institute for Human Genetics, Sioux Falls, SD, USA

[5]Department of Internal Medicine, Sanford School of Medicine, University of South Dakota, Sioux Falls, SD, USA

[6]Department of Psychiatry, Sanford School of Medicine, University of South Dakota, Sioux Falls, SD, USA

*** Corresponding author**

E-mail: bburrell@usd.edu (BDB)

Abstract

The medicinal leech, *Hirudo verbana*, is a powerful model organism for investigating fundamental neurobehavioral processes. The well-documented arrangement and properties of *H. verbana*'s nervous system allows changes at the level of specific neurons or synapses to be linked to physiological and behavioral phenomena. Juxtaposed to the extensive knowledge of *H. verbana's* nervous system is a limited, but recently expanding, portfolio of molecular and multi-omics tools. Together, the advancement of genetic databases for *H. verbana* will complement existing pharmacological and electrophysiological data by affording targeted manipulation and analysis of gene expression in neural pathways of interest. Here, we present the first draft genome assembly for *H. verbana*, which is approximately 250 Mbp in size and consists of 61,282 contigs. Whole genome sequencing was conducted using an Illumina sequencing platform followed by genome assembly with CLC-Bio Genomics Workbench and subsequent functional annotation. Ultimately, the diversity of organisms for which we have genomic information should parallel the availability of next generation sequencing technologies to widen the comparative approach to understand the involvement and discovery of genes in evolutionarily conserved processes. Results of this work hope to facilitate comparative studies with *H. verbana* and provide the foundation for future, more complete, genome assemblies of the leech.

Introduction

The medicinal leech (*Hirudo verbana*) has been repurposed from an ancient bloodletting instrument [1] to a widely utilized invertebrate model system in biomedical research [2]. The well-documented and accessible central nervous system of the leech allows for precise selection of neurons for electrophysiological studies based on their characteristic morphologies, positioning, and biophysical properties [3, 4]. Fundamental discoveries have been made using *Hirudo* in a variety of disciplines that include central pattern generators, behavioral choice, learning and memory, synaptic signaling, neuroethology, neuro-injury and repair, and neurodevelopment [5-9]. Extensive research has also been devoted to examining the proteins secreted during leech hematophagy, which has longstanding applications in inflammation and coagulation [10]. Genomic insights into the medicinal leech will facilitate a more comprehensive approach into the evolutionary conservation of genes involved in the mechanistic processes that the medicinal leech has been used to help elucidate.

Despite these well-documented advantages of the medicinal leech for addressing various research questions, the leech lacks the molecular and genetic tools in comparison to alternative model organisms [11, 12]. For example, *Caenorhabditis elegans* and *Drosophila melanogaster* have extensive resources for targeted genome engineering in addition to optogenetic tools for electrophysiology and behavior manipulation [13-16]. Improving the genomic resources of organisms like the medicinal leech will promote more inclusive comparative genomics approaches to identifying conserved structural and functional gene signatures involved in human health and disease. Moreover, for many years, the medicinal leech community had been inadvertently aggregating four species of medicinal leeches: *H. medicinalis*, *H. verbana*, *H.*

56 *orientalis*, and *H. troctina* [17-20]. This misunderstanding regarding the taxonomic classification
57 of these leech subspecies has led to some confusion surrounding appropriate cataloging of
58 preliminary leech omics databases [21, 22]. Finally, in spite of the advancements in sequencing
59 technology, most of the existing sequence repositories for the medicinal leeches have been
60 comprised of expressed sequence tag [23] and transcriptomic databases [24-26], with recent
61 leech genomics work centering around *H. medicinalis* despite the prominence of *H. verbana* in
62 neuroscience research [27, 28]. Moreover, other leech genomes are also becoming available,
63 including that for the Asian Buffalo leech, *Hirudinaria manillensis* [29]. This work presents the
64 first draft genome for *H. verbana,* which consists of 250 Mbp, 61,282 contigs, an N50 of 8,638
65 bp, and a GC content of 38%. This draft genome, in addition to the growing transcriptomic
66 resources for *H. verbana* [24, 26, 30] and the aforementioned genomic databases for the closely
67 related leech species, will help accelerate studies seeking to link the molecular basis of previous
68 and ongoing functional studies utilizing medicinal leeches.

69 # Materials and Methods

70 Tissue collection and DNA extraction: High molecular weight genomic DNA was isolated from
71 muscle of three specimens of *H. verbana* (obtained from Niagara Leeches, Cheyenne, WY) in
72 separate preparations using the QIAamp DNA Mini Kit Cat. No. 51304 (Qiagen; Hilden,
73 Germany). In accordance with the protocol, 25 mg of dissected muscle tissue from the body wall
74 was used as input for each of the DNA extractions. The DNA from the three different isolations
75 was pooled, concentrated with Ampure XP beads (Beckman Coulter; Brea, CA) and 500 ng (10
76 ng/µL) was utilized for sequencing library preparation.
77
78 Library preparation: Sequencing libraries were prepared using the TruSeq Synthetic Long-Read
79 DNA Library Prep Kit (Illumina, Inc; San Diego CA). Three sequencing libraries were prepared
80 by loading the sample into a Covaris g-TUBE (part no. 520079), and centrifuging twice at 4200
81 x g for 1 minute. Briefly, the DNA was fragmented to approximately 8-10 kb and ligated with
82 adapters, which mark the end of contigs during data analysis. Following a dilution to limit the
83 number of DNA molecules in each well of a 384-well plate, long-range PCR was performed to
84 enrich for DNA fragments with appropriate adapters. The DNA in each well was treated with
85 the Nextera transposome, which fragments and simultaneously adds adapters to DNA. Indexing-
86 PCR was used to barcode the DNA in each well of the 384-well plate. The resulting products
87 were pooled and bead size-selection was performed. The average size of the final libraries was
88 ~725 bp as measured with a High Sensitivity DNA chip on a 2100 Bioanalyzer (Agilent; Santa
89 Clara, CA). The concentration of each library was determined by quantitative PCR (qPCR) via
90 the KAPA Library Quantification Kit for Next Generation Sequencing (KAPA Biosystems;
91 Woburn, MA).

92 Whole genome sequencing: Libraries were normalized to 2 nmol/L in 10 mM Tris-Cl, pH 8.5
93 with 0.1% Tween 20. Prior to cluster amplification, the libraries were denatured with 0.05 N
94 NaOH and diluted to 20 pmol/L. Paired-end cluster generation of denatured templates was
95 performed according to the manufacturer's instructions (Illumina, San Diego, CA) utilizing the
96 HiSeq Rapid PE Cluster Kit v2 chemistry and flow cells. Libraries were optimally clustered at
97 11 pmol/L with a 1% PhiX spike-in. Sequencing-by-synthesis was performed on a HiSeq 2500
98 utilizing v2 chemistry with paired-end 101 bp reads and an 8 bp index read.
99
100 Long read and genome assembly: A total of 1,862,297,140 bp of 2 x 101 bp reads were obtained
101 from three flow cells. Sequence read data were processed and converted to short-read FASTQ
102 format by Illumina BaseSpace analysis software (v2.0.13). The short reads from each plate were
103 individually processed in three runs to construct primary contigs using the TruSPAdes assembly
104 software (v1.1.0) [31], which were merged into one assembly using CLC-Bio Genomics
105 Workbench *De Novo* Assembly (Qiagen, v11.0.1) with default parameters. Thorough quality
106 control was performed on the raw short read data using FastQC
107 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the Phred score, presence
108 of repeat reads, non-nucleotide content, GC content, and duplicated read contents. The quality of
109 the primary contigs assembled by the TruSPAdes algorithm were assessed by Quast [32].
110
111 Repeat sequence and BUSCO annotation: Repeatmasker [33] was used to annotate repeating
112 sequences and transposable elements in the *H. verbana* genome assembly. Repeatmasker was
113 configured with the pooled databases RepBase-2017 [34] and Dfam-2017 consensus databases
114 [35], and Tandem Repeats Finder, ran with RMBlast (v2.2.27) search engine. Additionally, the

115 completeness and quality of the draft genome was evaluated using a BUSCO (Benchmarking
116 Universal Single-Copy Orthologs) assessment that matched our newly assembled sequences to
117 the metazoan OrthoDB v9.1 [36-38]. BUSCO.v3 was configured using AUGUSTUS gene
118 predictor [39], HMMER [40, 41], and NCBI-BLAST+ [42].
119
120 Functional annotation and orthologous analysis: In order to elucidate functional annotation and
121 gene ontology annotation, NCBI Blast+, UniprotKB [43], and the Blast2GO software suite [44]
122 integrated with InterProScan [45] were implemented. Furthermore, through locally constructed
123 databases in CLC-Bio Genomics Workbench, we utilized NCBI Blastn [46] on our genome
124 against closely related databases for the following closely related polychaete annelids: *H.*
125 *medicinalis* (GenBank: EY478949-EY505781) , *Helobdella robusta* (GCA_000326865.1), and
126 *Capitella teleta* (GCA_000328365.1).
127
128 Gene prediction and macro-synteny analysis: Gene predictions were performed using the
129 MAKER2 [47] genome annotation pipeline with SNAP [48] against the nematode
130 *Caenorhabditis elegans* and two other annelids: *C. teleta*, and *H. robusta*. The *ab-initio* gene
131 predictor, SNAP, was trained three times to improve performance. The *H. verbana* draft genome
132 assembly was aligned to the *C. elegans* genome and scaffolds of *H. robusta*. Circos [49] was
133 used to generate the circular genome alignment figures to analyze the anchoring of the top 600
134 *H. verbana* contigs onto the six chromosomes of *C. elegans* [50]. Additonally, the draft genome
135 top 600 contigs were mapped and anchored to the 1,991 scaffolds of the *H. robusta* genome.
136

137 Phylogenetic reconstruction: OrthoFinder2 [51] was utilized for comparative genomics between
138 our draft genome for *H. verbana*, and the protein sequence databases for six other organisms: *H.*
139 *medicinalis*, *H. robusta*, *C. teleta*, *C. elegans* (GCA_000002985.3), and the chordates *Mus*
140 *musculus* (GCA_000001635.8) and *Homo sapiens* (GCA_000001405.27). Orthofinder was
141 configured with the DIAMOND search engine [52], MCL clustering algorithm [53], and
142 FastTree [54] to construct the rooted phylogenetic tree which was visualized with Phylo.io [55].
143

144 Species Identification: The presence of mitochondrial contigs was assessed by comparing the
145 draft mitochondrial genomes for *H. medicinalis* and *H. verbana* [56] to our draft genome. NCBI
146 Blastn [46] was used on our draft genome against the mitochondrial DNA sequences
147 downloaded from NCBI GenBank. The phylogenetic relationships between the mitochondrial
148 contigs present in our draft genome, *H. medicinalis*, *H. verbana*, as well as other Hirudinea,
149 *Whitmania pigra* [57] and *Hirudo nipponia* [58] via their mitochondrial genomes was also
150 analyzed. The evolutionary analysis was performed in MEGA X (v11.0) [59, 60] using Multiple
151 Sequence Comparison by Log- Expectation (MUSCLE) [61], and the phylogenetic tree was
152 constructed using the Maximum Likelihood method with the Tamura-Nei model [62] and node
153 confidence was assessed with a 500 bootstrap trials.

154 # Results and Discussion

155 Next-generation sequencing leveraging an Illumina HiSeq-2500 platform was employed to
156 construct the first draft genome for *H. verbana*, the medicinal leech. A total of 188 Gbp were
157 generated that encompassed 1,862,297,140 bp of 101bp x 2 paired short reads (Table 1A).
158 Quality control of the raw short reads performed by FastQC and MultiQC [63] revealed that the
159 data had an average phred score >30 (S1 Appendix). The short reads were barcode assembled
160 individually for each plate using TruSPAdes assembler software into TruSeq synthetic long
161 reads. The long reads (Table 1B) were 190,514 bp, 198,741 bp, and 193,658 bp for each plate,
162 respectively, and had an average N50 of 7623 bp. Together, these synthetic long reads consisted
163 of 582,913 sequences, 3,429,493,670 bp, and had an estimated coverage of 6.9X. Quast
164 assessment of both the synthetic long reads and final assembly maintained a phred score of 30.
165 The draft genome was constructed from the TruSeq synthetic long reads using CLC-Bio
166 Genomics Workbench to produce the draft genome assembly for *H. verbana*. Prior to arriving at
167 the final assembly that used a combination of TruSPAdes synthetic long reads and CLC-Bio
168 Genomics Workbench, we performed a thorough assessment of multiple assemblers using long
169 reads formed both by Illumina BaseSpace analysis software and TruSPAdes in conjunction with
170 SOAPdenovo2 [64], Megahit [65], Spades [66], Ray [67], and Velvet [68]. Ultimately, we are
171 reporting the TruSPAdes and CLC-Bio assembly approach because it had the most coverage and
172 performed best under downstream assessment described below.

173

174

175

176 **Table 1: Statistics of the *de novo* draft genome assembly for *Hirudo verbana***

(A) Genome sequencing reads obtained from Illumina HiSeq via Basespace			
	Reads (Single, 101bp)	Reads (Pair, 101bp x 2)	Bases (Pair)
Plate 1	266,688,108	533,376,216	53,870,997,816
Plate 2	328,593,254	657,186,508	66,375,837,308
Plate 3	335,867,208	671,734,416	67,845,176,016
Total reads	931,148,570	1,862,297,140	188,092,011,140
(B) TruSPAdes Truseq Synthetic Long-Read Assembly Statistics			
	Assembled reads	Total bases	N50
Plate 1	190,514	1,117,298,449	7,612
Plate 2	198,741	1,172,622,590	7,634
Plate 3	193,658	1,139,572,631	7,624
(C) Summary details of *Hirudo verbana* draft genome assembly			
Estimated genome size		250 Mb	
Total assembly length		250 Mb	
Total sequences		582,913	
Short read coverage		627X	
Long read coverage		6.9X	
Contigs		61,282	
N75		4807 bp	
N50		8638 bp	
N25		14800 bp	
Minimum contig		200 bp	
Maximum contig		154993 bp	
Average		4084 bp	
Nucleotide frequency			
Adenine (A)		30.90%	
Cytosine (C)		19.10%	
Guanine (G)		19.00%	
Thymine (T)		31.00%	

177 (A) Whole genome sequencing reads obtained from Illumina HiSeq for H. verbana de novo draft
178 genome assembly (B) Statistics for barcode-assembled synthetic long reads generated using
179 TruSPAdes (C) Summary of assembly statistics of the draft genome for H. verbana.

180

181 The final draft genome assembly presented here is 250,270,938 bp in size, which is
182 comparable to the genome assembly statistics for other annelids *C. teleta* (240 Mbp) and *H.*
183 *robusta* (310 Mbp). The genome assembly consists of 61,282 contigs that have a minimum
184 length of 200 bp, a maximum length of 154,993 bp, and an N50 of 8,638 bp (Table 1C). N50 is
185 important in an assembly assessment as it measures contigs lengths distribution and informs on
186 the internal parameter of the assembly. However, it does not inform or guarantee the
187 goodness/completeness of an assembly [69]. For example, filtering out of a shortest contigs or
188 the scaffolding that links contigs could increase the N50 without improving the assembly
189 completeness. Also, when analyzing the N50, a comparison should be made between draft
190 genomes of a similar length, or it will be misleading [70]. In this assembly, we picked the
191 assembly run with 200 bp parameter setting as the shortest contig cutoff. While assemblies were
192 tested with higher cutoffs, specifically 500 bp and 1000 bp, the chosen setting produced the best
193 outcome in in terms of balance of genome size coverage and N50 criteria. A condensed summary
194 comparing our draft genome to the 10 available annelid genomes is also provided in S1
195 Table.[27-29, 71-76] For preliminary validation of the quality of the draft genome assembly, the
196 61,282 contigs were mapped back to the raw short reads using the mapping module in CLC-Bio
197 Genomics Workbench. The result demonstrated that 86.72% of the assembly mapped back onto
198 the short reads, leaving 13.28% unmapped. Among the contigs that were reported to map back,
199 85.77% had identical base pair matching.

200 Next, the repeating segments and transposable elements of the draft genome were
201 annotated using RepeatMasker. An estimated 6.67% of the genome assembly (16,685,142 bp of
202 the total 250 Mbp) was determined to be repetitive or transposable elements, with a majority
203 consisting of simple repeats, interspersed repeats, and low complexity repeats (S2 Table).

204 Moreover, the completeness of the draft genome was assessed with a BUSCO analysis for the
205 presence of metazoan-specific orthologues. The metazoan BUSCO that we implemented
206 consisted of 978 genes, and our assembly returned 809 (82.70%) as complete, 533 (54.50%) as
207 complete and single-copy, 276 (28.20%) complete and duplicated, 70 (7.20%) fragmented, and
208 99 (10.10%) as missing (Table 2A). Overall, our genome has a completeness score of 89.9%
209 (82.70% complete + 7.20% fragmented).

210 **Table 2. Analysis of the completeness and orthology of the *H. verbana* genome.**

(A) BUSCO statistics of genome based on 978 metazoan-conserved genes		
BUSCO	Genes Present	Percentage (%)
Complete BUSCOs (C)	809	82.70%
Complete and Single-copy BUSCOs (S)	533	54.50%
Complete and duplicated BUSCOs (D)	276	28.20%
Fragmented BUSCOs (F)	70	7.20%
Missing BUSCOs (M)	99	10.10%
(B) Genome functional annotation		
	Number of contigs	Percentage (%)
Total	61,282	
Annotated		
InterProScan	5,000	8.16%
GO	2,800	4.57%
UniProtKB	44,000	71.80%
Unannotated	9,482	15.47%
(C) Comparison of gene prediction and annotation to related organisms		
Species	Size of genome assembly (Mbp)	Predicted number of genes
Lottia gigantea	348	23,800
Capitella teleta	324	32,389
Helobdella robusta	228	23,400
Hirudo verbana	250	26,210

211 (A) BUSCO statistics assessing the completeness of the H. verbana draft genome based on 978
212 metazoan-conserved genes (B) Summary of structural functional annotation for H. verbana draft
213 genome (C) Comparison of H. verbana draft genome size and predicted number of genes to 3
214 closely related spiralian genomes (two annelids and one mollusc).

215 The sequence homology and similarity of the draft genome for *H. verbana* was assessed
216 by NCBI BLAST+ against the genomic and transcriptomic sequences available for *H.*
217 *medicinalis*, *H. robusta*, and *C. teleta*. Approximately 94% of the draft genome sequences had an
218 identity match within the queried databases, 5.5% exhibited at least 70% similarity, and the
219 remaining 0.5% was unidentified. Functional annotation was assessed using fast-BlastX [77]
220 against the Animalia NCBI Refseq database in Blast2GO [78]. From total assembly, 1,178
221 contigs returned significant blast hits at an e-value threshold of 10^{-10}. The top 20 gene ontology
222 (GO) terms [79] for each classification – biological process (BP), molecular function (MF), and
223 cellular component (CC) – at GO level 5 are displayed in Fig 1. Moreover, the draft genome was
224 aligned to the genome of *C. elegans*. The draft genome contigs for *H. verbana* were mapped and
225 anchored to the 6 chromosomes of *C. elegans* (Fig 2A) and to the *H. robusta* scaffolds at
226 position 15000-17000 for contig *hv*7 on scaffold *he49* (Fig 2B). A majority of the mapped
227 sequences achieved better fit onto chromosomes 1 and X of the *C. elegans* genome (S3 Table).

228 Using OrthoFinder, orthologues were generated from gene families of our draft genome
229 for *H. verbana*, *H. medicinalis*, *H. robusta*, C. *teleta*, *C. elegans*, *M. musculus*, and *H. sapiens*.
230 The phylogenetic tree was reconstructed using OrthoFinder after it identified the highest
231 similarity content between the draft genome and the reference organisms. The phylogenetic tree
232 (Fig 3A) appropriately placed *H. verbana* adjacent to its closest relative, *H. medicinalis*,
233 demonstrating their last known divergence in genus *Hirudo*. Predictive protein-coding gene
234 sequences were identified based on conserved protein signatures and domains with UniprotKB,
235 Blast2Go, and InterProScan. A total of 84.53% of the contigs were annotated for a protein-
236 coding function, of these, 8.16% were identified by InterProScan, 4.57% by Blast2GO, and
237 71.80% by UniProtKB (Table 2B). Lastly, two-pass annotation in MAKER predicted 26,210

238 protein-coding genes in the draft genome, which is similar to the first reports of draft genomes
239 for fellow lophotrochozoans *C. teleta*, *H. robusta*, and *L. gigantea,* a gastropod mollusc (Table
240 2C).

241 For further confidence that the leech used in this study were *H. verbana*, we searched our
242 draft genome for mitochondrial contigs using the draft mitochondrial genomes for *H. medicinalis*
243 and *H. verbana*. The results of the mitochondrial contig analysis suggest that the leeches used to
244 collect genomic DNA in our study were of the species *H. verbana*, which is emphasized by the
245 mitochondrial contigs found in our draft genome forming a clade with the *H. verbana* draft
246 mitochondrial genome in the phylogenetic tree (Fig 3B). The results of the mitochondrial
247 phylogenetic analysis, in addition to the characteristic coloration patterns observed on the dorsal
248 side of the leeches, indicate that the leeches used in this study were *H. verbana*.

249 This study is the first to publish an annotated draft genome sequence for the medicinal
250 leech, *H. verbana*. Overall, the genome assembly consists of 250 Mbp and 61,282 contigs,
251 84.53% of which have been predicted to contain a protein-coding function. The draft genome is
252 also predicted to contain 26,210 protein-coding genes and a repetitive content of 6.67%. The raw
253 short-read sequence data, synthetic long-reads, and assembled contigs for the present study have
254 been deposited into NCBI accession number JAIQDV000000000. The draft genome assembly
255 will assist in providing tools to understand the underlying molecular processes involved in
256 ongoing studies in neurophysiology, developmental biology, and neuroethology that utilize *H.*
257 *verbana* [80-83]. Whole-genome characterization for Hirudinae *H. medicinalis* [27, 28], *H.*
258 *manillensis* [29] and *H. verbana* is expanding and will help distinguish and clarify distinct
259 genetic undertones of these previously amalgamated species. Future efforts to better annotate and
260 complete a genome for *H. verbana* will enable insight into genetic mechanisms of processes

investigated with this model organism [83-85], and advance more robust cross-species validation of comparative principles in next-generation biomedical research techniques.

Acknowledgements

This work was supported by a pilot grant from the University of South Dakota (USD) Center for Brain and Behavior Research (CBBRe), the USD Neuroscience, Nanotechnology, and Networks Program (DGE-1633213), and the National Science Foundation Graduate Research Fellowship (DGE-1545679). Authors acknowledge the use of computational resources of the USD HPC cluster and the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by the National Science Foundation grant number ACI-1548562.

References

1. Phillips, A.J. and M.E. Siddall, *Poly-paraphyly of Hirudinidae: many lineages of medicinal leeches.* BMC Evol Biol, 2009. **9**: p. 246.
2. Wagenaar, D.A., *A classic model animal in the 21st century: recent lessons from the leech nervous system.* J Exp Biol, 2015. **218**(Pt 21): p. 3353-9.
3. Nicholls, J.G. and D.A. Baylor, *Specific modalities and receptive fields of sensory neurons in CNS of the leech.* J Neurophysiol, 1968. **31**(5): p. 740-56.
4. Kristan, W.B., *Neuronal changes related to behavioral changes in chronically isolated segments of the medicinal leech.* Brain Research, 1979. **167**(1): p. 215-220.
5. Kristan, W.B., Jr., R.L. Calabrese, and W.O. Friesen, *Neuronal control of leech behavior.* Prog Neurobiol, 2005. **76**(5): p. 279-327.
6. Crisp, K.M. and B.D. Burrell, *Cellular and Behavioral Properties of Learning in Leech and Other Annelids* in *Annelids in Modern Biology*, D.H. Shain, Editor. 2009, John Wiley & Sons: New York.
7. Duan, Y., et al., *Repair and regeneration of functional synaptic connections: cellular and molecular interactions in the leech.* Cell Mol Neurobiol, 2005. **25**(2): p. 441-50.
8. Eisenhart, F.J., T.W. Cacciatore, and W.B. Kristan, Jr., *A central pattern generator underlies crawling in the medicinal leech.* J Comp Physiol A, 2000. **186**(7-8): p. 631-43.
9. Yuan, S. and B.D. Burrell, *Nonnociceptive afferent activity depresses nocifensive behavior and nociceptive synapses via an endocannabinoid-dependent mechanism.* J Neurophysiol, 2013. **110**(11): p. 2607-16.

292 10. Lemke, S., et al., *May salivary gland secretory proteins from hematophagous leeches*
293 *(Hirudo verbana) reach pharmacologically relevant concentrations in the vertebrate*
294 *host?* PLoS One, 2013. **8**(9): p. e73809.
295 11. Nobrega, M.A. and L.A. Pennacchio, *Comparative genomic analysis as a tool for*
296 *biological discovery.* J Physiol, 2004. **554**(Pt 1): p. 31-9.
297 12. Clarke, S.L., et al., *Human developmental enhancers conserved between deuterostomes*
298 *and protostomes.* PLoS Genet, 2012. **8**(8): p. e1002852.
299 13. Dickinson, D.J. and B. Goldstein, *CRISPR-Based Methods for Caenorhabditis elegans*
300 *Genome Engineering.* Genetics, 2016. **202**(3): p. 885-901.
301 14. Gratz, S.J., et al., *CRISPR-Cas9 Genome Editing in Drosophila.* Curr Protoc Mol Biol,
302 2015. **111**: p. 31 2 1-31 2 20.
303 15. Simpson, J.H. and L.L. Looger, *Functional Imaging and Optogenetics in Drosophila.*
304 Genetics, 2018. **208**(4): p. 1291-1309.
305 16. Leifer, A.M., et al., *Optogenetic manipulation of neural activity in freely moving*
306 *Caenorhabditis elegans.* Nat Methods, 2011. **8**(2): p. 147-52.
307 17. Kutschera, U. and J. Elliott, *The European medicinal leech Hirudo medicinalis L.:*
308 *Morphology and occurrence of an endangered species.* Zoosystematics and Evolution,
309 2014. **90**(2): p. 271-280.
310 18. Siddall, M.E., et al., *Diverse molecular data demonstrate that commercially available*
311 *medicinal leeches are not Hirudo medicinalis.* Proc Biol Sci, 2007. **274**(1617): p. 1481-7.
312 19. Petrauskiené, L., O. Utevska, and S. Utevsky, *Can different species of medicinal leeches*
313 *(Hirudospp.) interbreed?* Invertebrate Biology, 2009. **128**(4): p. 324-331.

314 20. Trontelj, P. and S.Y. Utevsky, *Phylogeny and phylogeography of medicinal leeches*
315 *(genus Hirudo): fast dispersal and shallow genetic structure.* Mol Phylogenet Evol, 2012.
316 **63**(2): p. 475-85.
317 21. Trontelj, P. and S.Y. Utevsky, *Celebrity with a neglected taxonomy: molecular*
318 *systematics of the medicinal leech (genus Hirudo).* Mol Phylogenet Evol, 2005. **34**(3): p.
319 616-24.
320 22. Bely, A.E. and D.A. Weisblat, *Lessons from leeches: a call for DNA barcoding in the*
321 *lab.* Evol Dev, 2006. **8**(6): p. 491-501.
322 23. Macagno, E.R., et al., *Construction of a medicinal leech transcriptome database and its*
323 *application to the identification of leech homologs of neural and innate immune genes.*
324 BMC Genomics, 2010. **11**: p. 407.
325 24. Northcutt, A.J., et al., *An annotated CNS transcriptome of the medicinal leech, Hirudo*
326 *verbana: De novo sequencing to characterize genes associated with nervous system*
327 *activity.* PLoS One, 2018. **13**(7): p. e0201206.
328 25. Hibsh, D., et al., *De novo transcriptome assembly databases for the central nervous*
329 *system of the medicinal leech.* Sci Data, 2015. **2**: p. 150015.
330 26. Heath-Heckman, E., et al., *Transcriptional profiling of identified neurons in leech.* BMC
331 Genomics, 2021. **22**(1): p. 215.
332 27. Babenko, V.V., et al., *Draft genome sequences of Hirudo medicinalis and salivary*
333 *transcriptome of three closely related medicinal leeches.* BMC Genomics, 2020. **21**(1): p.
334 331.

335 28. Kvist, S., et al., *Draft genome of the European medicinal leech Hirudo medicinalis*
336 *(Annelida, Clitellata, Hirudiniformes) with emphasis on anticoagulants.* Sci Rep, 2020.
337 **10**(1): p. 9885.
338 29. Guan, D.L., et al., *Draft Genome of the Asian Buffalo Leech Hirudinaria manillensis.*
339 Front Genet, 2019. **10**: p. 1321.
340 30. Stowasser, A., et al., *Electrophysiology and transcriptomics reveal two photoreceptor*
341 *classes and complex visual integration in Hirudo verbana.* J Exp Biol, 2019. **222**(Pt 15).
342 31. Bankevich, A. and P.A. Pevzner, *TruSPAdes: barcode assembly of TruSeq synthetic long*
343 *reads.* Nat Methods, 2016. **13**(3): p. 248-50.
344 32. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.*
345 Bioinformatics, 2013. **29**(8): p. 1072-5.
346 33. Tarailo-Graovac, M. and N. Chen, *Using RepeatMasker to identify repetitive elements in*
347 *genomic sequences.* Curr Protoc Bioinformatics, 2009. **Chapter 4**: p. Unit 4 10.
348 34. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet
349 Genome Res, 2005. **110**(1-4): p. 462-7.
350 35. Hubley, R., et al., *The Dfam database of repetitive DNA families.* Nucleic Acids Res,
351 2016. **44**(D1): p. D81-9.
352 36. Simao, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness*
353 *with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-2.
354 37. Waterhouse, R.M., et al., *BUSCO Applications from Quality Assessments to Gene*
355 *Prediction and Phylogenomics.* Mol Biol Evol, 2018. **35**(3): p. 543-548.

356 38. Zdobnov, E.M., et al., *OrthoDB v9.1: cataloging evolutionary and functional annotations*
357 *for animal, fungal, plant, archaeal, bacterial and viral orthologs.* Nucleic Acids Res,
358 2017. **45**(D1): p. D744-D749.
359 39. Stanke, M., A. Tzvetkova, and B. Morgenstern, *AUGUSTUS at EGASP: using EST,*
360 *protein and genomic alignments for improved gene prediction in the human genome.*
361 Genome Biol, 2006. **7 Suppl 1**: p. S11 1-8.
362 40. Potter, S.C., et al., *HMMER web server: 2018 update.* Nucleic Acids Res, 2018. **46**(W1):
363 p. W200-W204.
364 41. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence*
365 *similarity searching.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W29-37.
366 42. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009.
367 **10**: p. 421.
368 43. UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Res, 2019.
369 **47**(D1): p. D506-D515.
370 44. Gotz, S., et al., *High-throughput functional annotation and data mining with the*
371 *Blast2GO suite.* Nucleic Acids Res, 2008. **36**(10): p. 3420-35.
372 45. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.*
373 Bioinformatics, 2014. **30**(9): p. 1236-40.
374 46. Johnson, M., et al., *NCBI BLAST: a better web interface.* Nucleic Acids Res, 2008.
375 **36**(Web Server issue): p. W5-9.
376 47. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database*
377 *management tool for second-generation genome projects.* BMC Bioinformatics, 2011.
378 **12**: p. 491.

379 48. Korf, I., *Gene finding in novel genomes.* BMC Bioinformatics, 2004. **5**: p. 59.
380 49. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.*
381 Genome Res, 2009. **19**(9): p. 1639-45.
382 50. Hillier, L.W., et al., *Genomics in C. elegans: so many genes, such a little worm.* Genome
383 Res, 2005. **15**(12): p. 1651-60.
384 51. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for*
385 *comparative genomics.* Genome Biol, 2019. **20**(1): p. 238.
386 52. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using*
387 *DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
388 53. Enright, A.J., S. Van Dongen, and C.A. Ouzounis, *An efficient algorithm for large-scale*
389 *detection of protein families.* Nucleic Acids Res, 2002. **30**(7): p. 1575-84.
390 54. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood*
391 *trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
392 55. Robinson, O., D. Dylus, and C. Dessimoz, *Phylo.io: Interactive Viewing and Comparison*
393 *of Large Phylogenetic Trees on the Web.* Mol Biol Evol, 2016. **33**(8): p. 2163-6.
394 56. Nikitina, A., et al., *Draft mitochondrial genomes of Hirudo medicinalis and Hirudo*
395 *verbana (Annelida, Hirudinea).* Mitochondrial DNA B Resour, 2016. **1**(1): p. 254-256.
396 57. Shen, X., et al., *The complete mitochondrial genome sequence of Whitmania pigra*
397 *(Annelida, Hirudinea): the first representative from the class Hirudinea.* Comp Biochem
398 Physiol Part D Genomics Proteomics, 2011. **6**(2): p. 133-8.
399 58. Xu, Y., et al., *Complete mitochondrial genome of Hirudo nipponia (Annelida,*
400 *Hirudinea).* Mitochondrial DNA A DNA Mapp Seq Anal, 2016. **27**(1): p. 257-8.

401 59. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing*
402 *Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
403 60. Stecher, G., K. Tamura, and S. Kumar, *Molecular Evolutionary Genetics Analysis*
404 *(MEGA) for macOS.* Mol Biol Evol, 2020. **37**(4): p. 1237-1239.
405 61. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high*
406 *throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
407 62. Tamura, K., *Estimation of the number of nucleotide substitutions when there are strong*
408 *transition-transversion and G+C-content biases.* Mol Biol Evol, 1992. **9**(4): p. 678-87.
409 63. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a*
410 *single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
411 64. Luo, R., et al., *SOAPdenovo2: an empirically improved memory-efficient short-read de*
412 *novo assembler.* Gigascience, 2012. **1**(1): p. 18.
413 65. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex*
414 *metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p.
415 1674-6.
416 66. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to*
417 *single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
418 67. Boisvert, S., F. Laviolette, and J. Corbeil, *Ray: simultaneous assembly of reads from a*
419 *mix of high-throughput sequencing technologies.* J Comput Biol, 2010. **17**(11): p. 1519-
420 33.
421 68. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de*
422 *Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.

423 69. Narzisi, G. and B. Mishra, *Comparing De Novo Genome Assembly: The Long and Short*
424 *of It.* PLOS ONE, 2011. **6**(4): p. e19175.
425 70. Thrash, A., F. Hoffmann, and A. Perkins, *Toward a more holistic method of genome*
426 *assembly assessment.* BMC Bioinformatics, 2020. **21**(4): p. 249.
427 71. Simakov, O., et al., *Insights into bilaterian evolution from three spiralian genomes.*
428 Nature, 2013. **493**(7433): p. 526-31.
429 72. Martin-Duran, J.M., et al., *Conservative route to genome compaction in a miniature*
430 *annelid.* Nat Ecol Evol, 2021. **5**(2): p. 231-242.
431 73. Shikuma, N.J., et al., *Stepwise metamorphosis of the tubeworm Hydroides elegans is*
432 *mediated by a bacterial inducer and MAPK signaling.* Proc Natl Acad Sci U S A, 2016.
433 **113**(36): p. 10097-102.
434 74. Cunha, L., et al., *A large set of microsatellites for the highly invasive earthworm*
435 *Amynthas corticis predicted from low coverage genomes.* Applied Soil Ecology, 2017.
436 **119**: p. 152-155.
437 75. Li, Y., et al., *Genomic adaptations to chemosymbiosis in the deep-sea seep-dwelling*
438 *tubeworm Lamellibrachia luymesi.* BMC Biol, 2019. **17**(1): p. 91.
439 76. Bhambri, A., et al., *Large scale changes in the transcriptome of Eisenia fetida during*
440 *regeneration.* PLoS One, 2018. **13**(9): p. e0204234.
441 77. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastBLAST: homology relationships for millions*
442 *of proteins.* PLoS One, 2008. **3**(10): p. e3589.
443 78. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status,*
444 *taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p.
445 D733-45.

446 79. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.*
447 Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.
448 80. Harley, C.M. and D.A. Wagenaar, *Scanning behavior in the medicinal leech Hirudo*
449 *verbana.* PLoS One, 2014. **9**(1): p. e86120.
450 81. Yazdani, N., et al., *Expression of a dominant negative mutant innexin in identified*
451 *neurons and glial cells reveals selective interactions among gap junctional proteins.* Dev
452 Neurobiol, 2013. **73**(8): p. 571-86.
453 82. Yuan, S. and B.D. Burrell, *Endocannabinoid-dependent long-term depression in a*
454 *nociceptive synapse requires coordinated presynaptic and postsynaptic transcription and*
455 *translation.* J Neurosci, 2013. **33**(10): p. 4349-58.
456 83. Baranzini, N., et al., *Antimicrobial Role of RNASET2 Protein During Innate Immune*
457 *Response in the Medicinal Leech Hirudo verbana.* Front Immunol, 2020. **11**: p. 370.
458 84. Grey, K.B., B.L. Moss, and B.D. Burrell, *Molecular identification and expression of the*
459 *NMDA receptor NR1 subunit in the leech.* Invert Neurosci, 2009. **9**(1): p. 11-20.
460 85. Kabeiseman, E., R. Paulsen, and B.D. Burrell, *Characterization of a monoacylglycerol*
461 *lipase in the medicinal leech, Hirudo verbana.* Comp Biochem Physiol B Biochem Mol
462 Biol, 2020. **243-244**: p. 110433.

463

464

465 # Figure Legends

466 **Figure 1. *H. verbana* draft genome gene ontology distribution.** Displayed are the number of
467 sequences present corresponding to the top 20 most frequent annotations for each classification.
468 Gene ontologies shown are grouped into the major categories of: biological process (BP, green),
469 molecular function (MF, blue), and cellular component (CC, yellow) at a gene ontology level of
470 5.

471

472 **Figure 2. Alignment of *H. verbana* draft genome to *C. elegans* and *H. robusta*.** (A) Alignment
473 (>100 bp) of top 600 *H. verbana* draft genome contigs to the chromosomes (I-X) of *C. elegans*.
474 (B) Alignment (>100 bp) of the top 600 *H. verbana* draft genome contigs to the scaffolds of the
475 *H. robusta* genome assembly. *H. verbana* contigs are in black and abbreviated as "Hv," the *C.*
476 *elegans* chromosomes are abbreviated "ce," and the *H. robusta* scaffolds are abbreviated "he."
477 Both the corresponding chromosomes and scaffolds are colored, and the projections are anchored
478 to different contigs from the draft genome.

479

480 **Figure 3. Genomic and mitochondrial phylogenetic relationships.** (A) Reconstructed
481 phylogenic tree based on orthologous gene families. (B) Phylogeny of our draft genome in
482 relation to 4 leech species by comparing the presence and similarity of mitochondrial contigs
483 through the Maximum Likelihood method via the Tamura 3-parameter model. NCBI accession
484 numbers are provided next to each given taxa name.

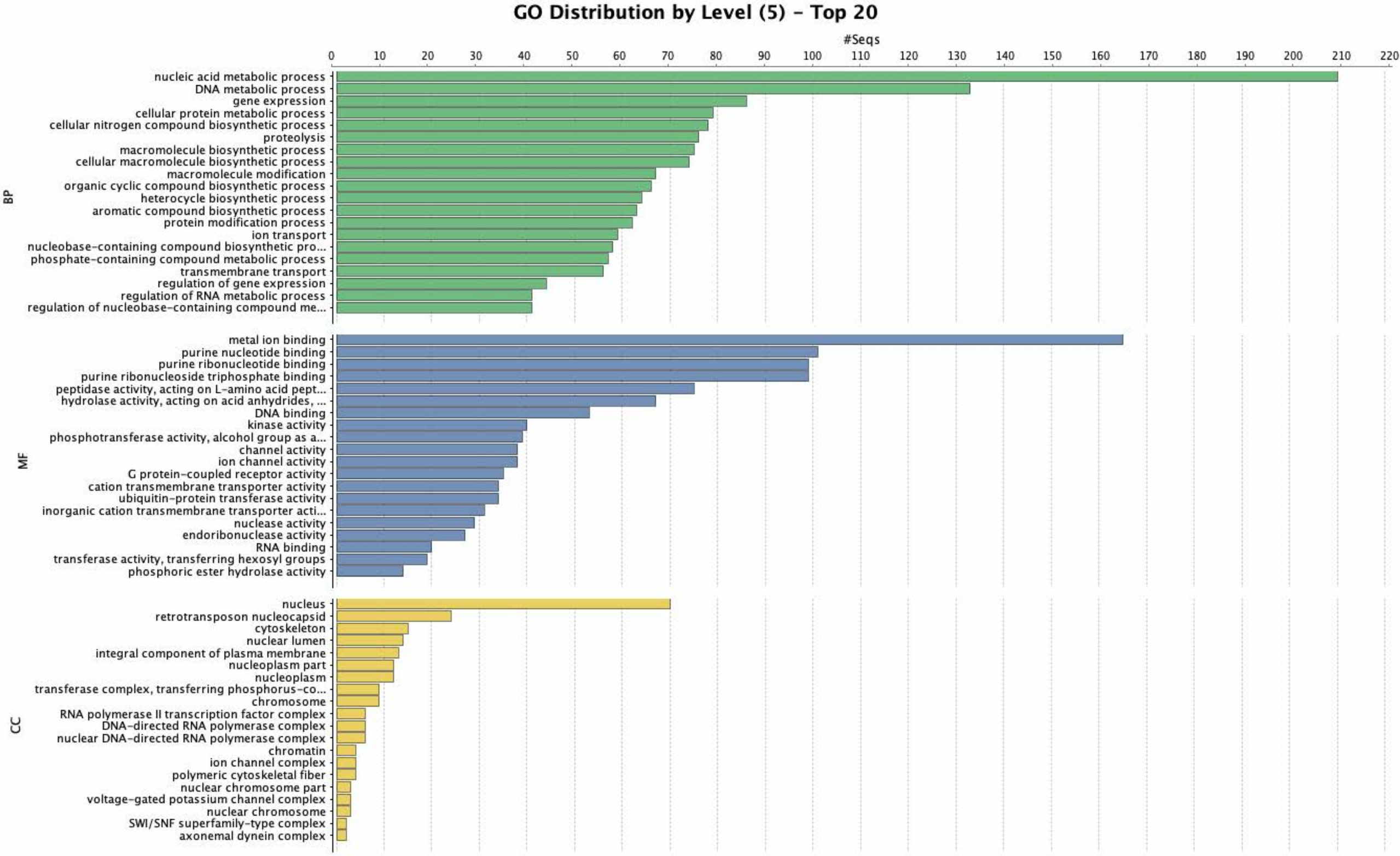
Supplementary Figure/Tables

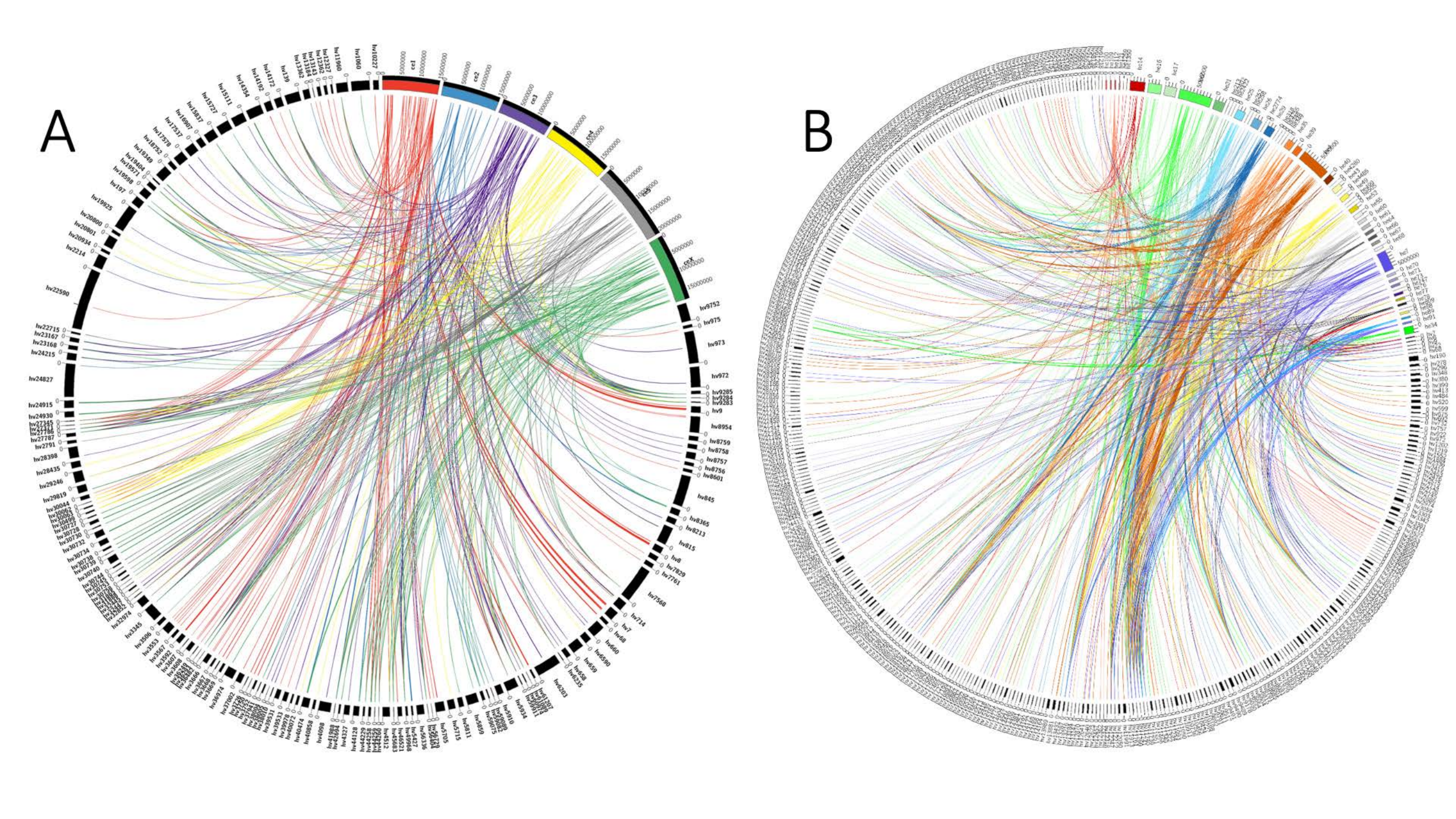
Supplementary Figure 1. Quality Control from Sequencing to Final Assembly Stage. (A) Integrated quality assessment for short sequencing reads generated by FastQC and MultiQC (B) GC content for final assembly and (C) TruSPAdes synthetic long reads after barcode assembly.

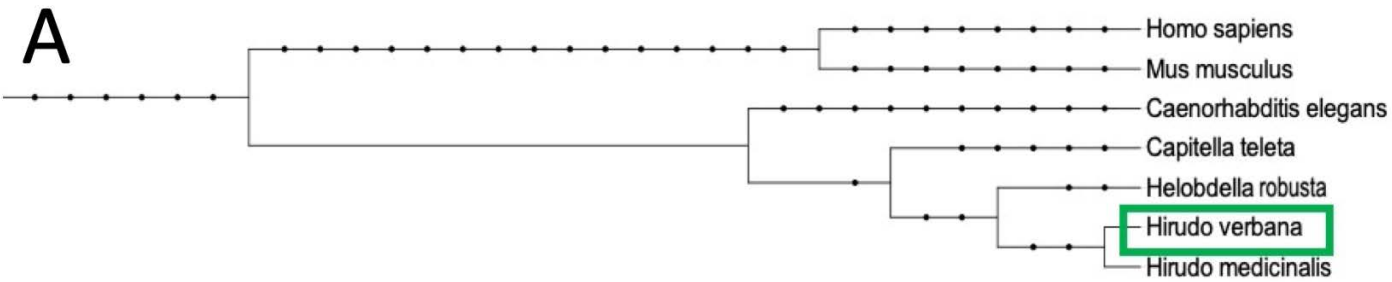
Supplementary Table 1. Annelid Genome Comparison. Comparison of select assembly statistics at the contig level (number of contigs, genome length, N50, and GC content) for annelid genomes. Accession numbers are provided, and the mode of sequencing technology and assembly software are also included for comparison.

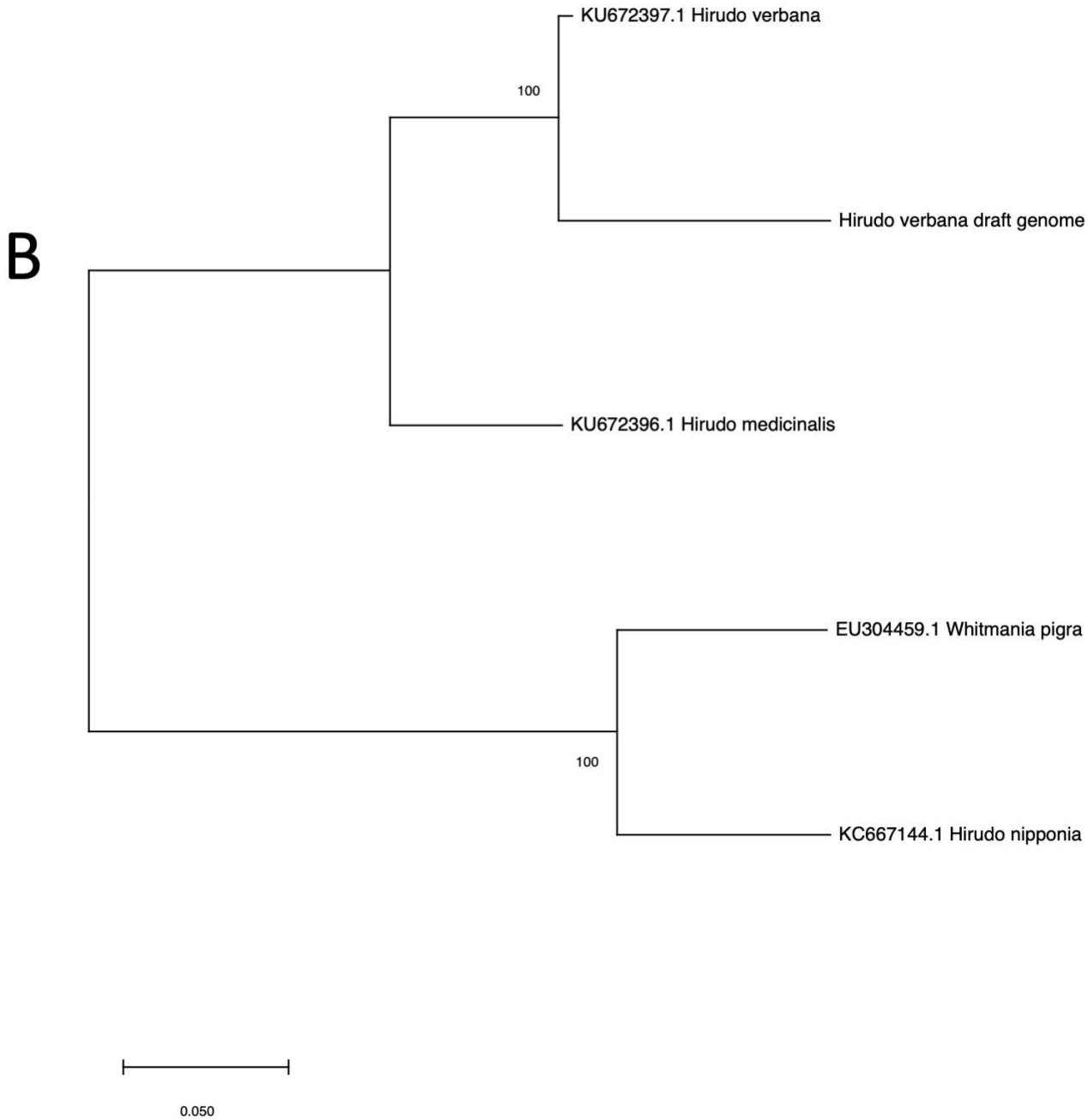
Supplementary Table 2. Repeat sequence and transposable element composition in the draft genome of *H. verbana*. Identification of repetitive elements in the *H. verbana* draft genome using RepeatMasker.

Supplementary Table 3. Chromosome rearrangement of *C. elegans* onto *H. verbana* contigs. Mapping of *H. verbana* draft genome contigs onto the chromosome scaffolds of *C. elegans.*


GO Distribution by Level (5) - Top 20
#Seqs
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
BP
nucleic acid metabolic process
DNA metabolic process
gene expression
cellular protein metabolic process
cellular nitrogen compound biosynthetic process
proteolysis
macromolecule biosynthetic process
cellular macromolecule biosynthetic process
macromolecule modification
organic cyclic compound biosynthetic process
heterocycle biosynthetic process
aromatic compound biosynthetic process
protein modification process
ion transport
nucleobase-containing compound biosynthetic pro...
phosphate-containing compound metabolic process
transmembrane transport
regulation of gene expression
regulation of RNA metabolic process
regulation of nucleobase-containing compound me...
MF
metal ion binding
purine nucleotide binding
purine ribonucleotide binding
purine ribonucleoside triphosphate binding
peptidase activity, acting on L-amino acid pept...
hydrolase activity, acting on acid anhydrides, ...
DNA binding
kinase activity
phosphotransferase activity, alcohol group as a...
channel activity
ion channel activity
G protein-coupled receptor activity
cation transmembrane transporter activity
ubiquitin-protein transferase activity
inorganic cation transmembrane transporter acti...
nuclease activity
endoribonuclease activity
RNA binding
transferase activity, transferring hexosyl groups
phosphoric ester hydrolase activity
CC
nucleus
retrotransposon nucleocapsid
cytoskeleton
nuclear lumen
integral component of plasma membrane
nucleoplasm part
nucleoplasm
transferase complex, transferring phosphorus-co...
chromosome
RNA polymerase II transcription factor complex
DNA-directed RNA polymerase complex
nuclear DNA-directed RNA polymerase complex
chromatin
ion channel complex
polymeric cytoskeletal fiber
nuclear chromosome part
voltage-gated potassium channel complex
nuclear chromosome
SWI/SNF superfamily-type complex
axonemal dynein complex


A
B


A
Homo sapiens
Mus musculus
Caenorhabditis elegans
Capitella teleta
Helobdella robusta
Hirudo verbana
Hirudo medicinalis


B
KU672397.1 Hirudo verbana
100
Hirudo verbana draft genome
KU672396.1 Hirudo medicinalis
EU304459.1 Whitmania pigra
100
KC667144.1 Hirudo nipponia
0.050

Organism	Accession Number	Contig Number	Total Genome Length (Mbp)	N50	GC content (%)	Sequencing Technology	Assembler
Dimorphilus gyrociliatus	GCA_904063045	779	77.892	979,161	31	PacBio	PBcR
Hirudo medicinalis	GCA_011800805	25,027	187.547	24,951	30	IonTorrent	SPAdes
Hirudo medicinalis	GCA_903470615	36,325	176.963	14,470	34.9	Illumina	Supernova
Hirudo verbana		**61,282**	**250.270**	**8,638**	**38**	**Illumina**	**CLCBio**
Hydroides elegans	GCA_001703475	370,534	1026.05	5,980	34.6	Illumina	Velvet
Amynthas corticis	GCA_900184025	3,225,964	1017.22	510	40.1	Illumina	SoapDenovo2
Lamellibrachia luymesi	GCA_009193005	55,092	687.706	24,086	35.1	Illumina	Platanus
Poecilobdella manillensis	GCA_015345955	467	151.673	2,277,518	36	PacBio	MECAT
Capitella teleta	GCA_000328365	49,393	333.283	21,930	41.8	Sanger	JGI Jazz
Helobdella robusta	GCA_000326865	12,292	235.376	52,195	34.2	Sanger	JGI Jazz
Eisenia fetida	GCA_003999395	536,508	1471.98	897	12.5	Illumina	CLCBio

Element classification		No. of elements	Length (bp)	% of sequence
Total length of contigs		61,282	250,270,938	
Repetitive sequences			16,685,142	6.67%
Transposable elements				
LTR		327	107,597	0.04%
	ERVL	36	8,360	0.00%
	ERVL-MaLRs	31	10,905	0.00%
	ERV_class I	30	3,225	0.00%
	ERV_class II	2	110	0.00%
DNA		870	132,616	0.05%
	hAT-Charlie	327	43,781	0.02%
	TcMar-Tigger	119	27085	0.01%
LINE		3,202	622,089	0.25%
	LINE1	112	35,062	0.01%
	LINE2	1,378	316,959	0.13%
	L3/CR1	1,411	237,912	0.10%
SINE		286	36,605	0.01%
	ALUs	93	20,833	0.01%
	MIRs	77	8,440	0.00%
Unclassified		76	5410	0.00%
Interspersed repeats			904317	0.36%
Small RNA		6	1261	0.00%
Satellites		43	3896	0.00%
Simple repeats		236935	15011804	6.00%
Low complexity		15265	765808	0.31%

C. elegans chromosome	**Hirudo verbana selected Contigs**
1	hv3592, hv3669, hv3750, hv5705, hv61207, hv68, hv7, hv8, hv9, hv3668, hv3752
2	hv5427, hv38001, hv18257, hv45683
3	hv6203, hv972, hv973, hv32447, hv38005, hv39978, hv15111, hv17537, hv2214, hv24215, hv23168, hv17578, hv714, hv29246, hv38010, hv60974, hv8213, hv8757, hv8758, hv8756, hv8759, hv4098, hv27371, hv27345
4	hv18257, hv714, hv29246, hv27371, hv27345
5	hv815, hv14354, hv19349, hv24915, hv24930, hv30728, hv30732, hv30740, hv30739, hv30727, hv30744, hv30745, hv18257, hv22715, hv29819, hv2791, hv30738, hv30753, hv30730, hv30758, hv30734, hv42694
X	hv14354, hv19349, hv24930, hv30728, hv30732, hv30740, hv30739, hv30727, hv18257, hv29819, hv42694, hv36974, hv5811, hv5859, hv30738, hv30753, hv30758, hv30734, hv30730, hv24215, hv23168, hv23167