1 **Title**

2 The Draft Reference Genome for *Hirudo verbana*, the Medicinal Leech

3 **Authors and Affiliations**

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

5 Etienne Gnimpieba[2,3], and Brian D. Burrell[1,2*]

6 [1]Division of Basic Biomedical Sciences and Center for Brain and Behavior Research, Sanford

7 School of Medicine, University of South Dakota, Vermillion, SD, USA

8 [2]Center for Brain and Behavior Research, University of South Dakota, SD, USA

9 [3]Biomedical Engineering Program, University of South Dakota, Sioux Falls, SD, USA

10 [4]Avera Institute for Human Genetics, Sioux Falls, SD, USA

11 [5]Department of Internal Medicine, Sanford School of Medicine, University of South Dakota,

12 Sioux Falls, SD, USA

13 [6]Department of Psychiatry, Sanford School of Medicine, University of South Dakota, Sioux

14 Falls, SD, USA

15 *** Corresponding author**

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

17

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. orientalis*, and *H. troctina* [17-20]. This misunderstanding regarding the taxonomic classification

of these leech subspecies has led to some confusion surrounding appropriate cataloging of preliminary leech omics databases [21, 22]. Finally, in spite of the advancements in sequencing technology, most of the existing sequence repositories for the medicinal leeches have been comprised of expressed sequence tag [23] and transcriptomic databases [24], with many centering around *H. medicinalis* despite the prominence of *H. verbana* in neuroscience research [25]. This work presents the first draft genome for *H. verbana,* which consists of 250 Mbp, 61,282 contigs, an N50 of 8,638 bp, and a GC content of 38%. This draft genome, in addition to the growing transcriptomic resources for *H. verbana* [26, 27] and draft genome assembly for *H. medicinalis* [28], will help accelerate studies seeking to link the molecular basis of previous and ongoing functional studies utilizing medicinal leeches.

Materials and Methods

Tissue collection and DNA extraction: High molecular weight genomic DNA was isolated from muscle of three specimens of *H. verbana* (obtained from Niagara Leeches, Cheyenne, WY) in separate preparations using the QIAamp DNA Mini Kit (Qiagen; Hilden, Germany). The DNA was pooled and 500 ng was utilized for sequencing library preparation.

Library preparation: Sequencing libraries were prepared using the TruSeq Synthetic Long-Read DNA Library Prep Kit (Illumina, Inc; San Diego CA). Three sequencing libraries were prepared following the manufacturer's recommendation. Briefly, the DNA was fragmented to approximately 8-10 kb and ligated with adapters, which mark the end of contigs during data analysis. Following a dilution to limit the number of DNA molecules in each well of a 384-well plate, long-range PCR was performed to enrich for DNA fragments with appropriate adapters. The DNA in each well was treated with the Nextera transposome, which fragments and simultaneously adds adapters to DNA. Indexing-PCR was used to barcode the DNA in each well

of the 384-well plate. The resulting products were pooled and bead size-selection was performed. The average size of the final libraries was ~725 bp as measured with a High Sensitivity DNA chip on a 2100 Bioanalyzer (Agilent; Santa Clara, CA). The concentration of each library was determined by quantitative PCR (qPCR) via the KAPA Library Quantification Kit for Next Generation Sequencing (KAPA Biosystems; Woburn, MA).

Whole genome sequencing: Libraries were normalized to 2 nmol/L in 10 mM Tris-Cl, pH 8.5 with 0.1% Tween 20. Prior to cluster amplification, the libraries were denatured with 0.05 N NaOH and diluted to 20 pmol/L. Paired-end cluster generation of denatured templates was performed according to the manufacturer's instructions (Illumina, San Diego, CA) utilizing the HiSeq Rapid PE Cluster Kit v2 chemistry and flow cells. Libraries were optimally clustered at 11 pmol/L with a 1% PhiX spike-in. Sequencing-by-synthesis was performed on a HiSeq 2500 utilizing v2 chemistry with paired-end 101 bp reads and an 8 bp index read.

Long read and genome assembly: A total of 1,862,297,140 bp of 2 x 101 bp reads were obtained from three flow cells. Sequence read data were processed and converted to short-read FASTQ format by Illumina BaseSpace analysis software (v2.0.13). The short reads from each plate were individually processed in three runs to construct primary contigs using the TruSPAdes assembly software (v1.1.0) [29], and were combined using CLC-Bio Genomics Workbench *De Novo* Assembly (Qiagen, v11.0.1). Thorough quality control was performed on the raw short read data using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) to assess the Phred score, presence of repeat reads, non-nucleotide content, GC content, and duplicated read contents. The quality of the primary contigs assembled by the TruSPAdes algorithm were assessed by Quast [30].

Repeat sequence and BUSCO annotation: Repeatmasker [31] was used to annotate repeating sequences and transposable elements in the *H. verbana* genome assembly. Repeatmasker was configured with the pooled databases RepBase [32] and Dfam-Consensus [33], RMBlast , and Tandem Repeats Finder. Additionally, the completeness and quality of the draft genome was evaluated using a BUSCO (Benchmarking Universal Single-Copy Orthologs) assessment that matched our newly assembled sequences to the metazoan OrthoDB v9.1 [34-36]. BUSCO.v3 was configured using AUGUSTUS gene predictor [37], HMMER [38, 39], and NCBI-BLAST+ [40].

Functional annotation and orthologous analysis: In order to elucidate functional annotation and gene ontology annotation, NCBI Blast+, UniprotKB [41], and the Blast2GO software suite [42] integrated with InterProScan [43] were implemented. Furthermore, through locally constructed databases in CLC-Bio Genomics Workbench, we utilized NCBI Blastn [44] on our genome against closely related databases for the following closely related polychaete annelids: *H. medicinalis* (GenBank: EY478949-EY505781) , *Helobdella robusta* (GCA_000326865.1), and *Capitella teleta* (GCA_000328365.1).

Gene prediction and macro-synteny analysis: Gene predictions were performed using the MAKER2 [45] genome annotation pipeline with SNAP [46] against the nematode *Caenorhabditis elegans* and two other annelids: *C. teleta*, and *H. robusta*. The *ab-initio* gene predictor, SNAP, was trained three times to improve performance. The *H. verbana* draft genome assembly was aligned to the *C. elegans* genomes. Circos [47] was used to generate the circular genome alignment figures to analyze the anchoring of the top 600 *H. verbana* contigs onto the six chromosomes of *C. elegans* [48].

Phylogenic reconstruction: OrthoFinder2 [49] was utilized for comparative genomics between our draft genome for *H. verbana*, and the protein sequence databases for six other organisms: *H. medicinalis*, *H. robusta*, *C. teleta*, *C. elegans* (GCA_000002985.3), and the chordates *Mus musculus* (GCA_000001635.8) and *Homo sapiens* (GCA_000001405.27). Orthofinder was configured with the DIAMOND search engine [50], MCL clustering algorithm [51], and FastTree [52] to construct the rooted phylogenic tree which was visualized with Phylo.io [53].

Results, Discussion, Conclusions

Next-generation sequencing leveraging an Illumina HiSeq-2500 platform was employed to construct the first draft genome for *H. verbana*, the medicinal leech. A total of 188 Gbp were generated that encompassed 1,862,297,140 bp of 101bp x 2 paired short reads (Table 1A). Quality control of the raw short reads performed by FastQC and MultiQC [54] revealed that the data had an average phred score >30 (S1 Appendix). The short reads were barcode assembled individually for each plate using TruSPAdes assembler software into TruSeq synthetic long reads. The long reads (Table 1B) were 190,514 bp, 198,741 bp, and 193,658 bp for each plate, respectively, and had an average N50 of 7623 bp. Together, these synthetic long reads consisted of 582,913 sequences, 3,429,493,670 bp, and had an estimated coverage of 6.9X. Quast assessment of both the synthetic long reads and final assembly maintained a phred score of 30. The draft genome was constructed from the TruSeq synthetic long reads using CLC-Bio Genomics Workbench to produce the draft genome assembly for *H. verbana*. Prior to arriving at the final assembly that used a combination of TruSPAdes synthetic long reads and CLC-Bio Genomics Workbench, we performed a thorough assessment of multiple assemblers using long reads formed both by Illumina BaseSpace analysis software and TruSPAdes in conjunction with

SOAPdenovo2 [55], Megahit [56], Spades [57], Ray [58], and Velvet [59]. Ultimately, we are reporting the TruSPAdes and CLC-Bio assembly approach because it had the most coverage and performed best under downstream assessment described below.

The final draft genome assembly presented here is 250,270,938 bp in size, which is comparable to the genome assembly statistics for other annelids *C. teleta* (240 Mbp) and *H. robusta* (310 Mbp). The genome assembly consists of 61,282 contigs that have a minimum length of 200 bp, a maximum length of 154,993 bp, and an N50 of 8,638 bp (Table 1C). For preliminary validation of the quality of the draft genome assembly, the 61,282 contigs were mapped back to the raw short reads using the mapping module in CLC-Bio Genomics Workbench. The result demonstrated that 86.72% of the assembly mapped back onto the short reads, leaving 13.28% unmapped. Among the contigs that were reported to map back, 85.77% had identical base pair matching.

Next, the repeating segments and transposable elements of the draft genome were annotated using RepeatMasker. An estimated 6.67% of the genome assembly (16,685,142 bp of the total 250 Mbp) was determined to be repetitive or transposable elements, with a majority consisting of simple repeats, interspersed repeats, and low complexity repeats (S1 Table). Moreover, the completeness of the draft genome was assessed with a BUSCO analysis for the presence of metazoan-specific orthologues. The metazoan BUSCO that we implemented consisted of 978 genes, and our assembly returned 809 (82.70%) as complete, 533 (54.50%) as complete and single-copy, 276 (28.20%) complete and duplicated, 70 (7.20%) fragmented, and 99 (10.10%) as missing (Table 2A). Overall, our genome has a completeness score of 89.9% (82.70% complete + 7.20% fragmented).

153 The sequence homology and similarity of the draft genome for *H. verbana* was assessed
154 by NCBI BLAST+ against the genomic and transcriptomic sequences available for *H.*
155 *medicinalis*, *H. robusta*, and *C. teleta*. Approximately 94% of the draft genome sequences had an
156 identity match within the queried databases, 5.5% exhibited at least 70% similarity, and the
157 remaining 0.5% was unidentified. Functional annotation was assessed using fast-BlastX [60]
158 against the Animalia NCBI Refseq database in Blast2GO [61]. From total assembly, 1,178
159 contigs returned significant blast hits at an e-value threshold of 10^{-10}. The top 20 gene ontology
160 (GO) terms [62] for each classification – biological process (BP), molecular function (MF), and
161 cellular component (CC) – at GO level 5 are displayed in Fig. 1. Moreover, the draft genome was
162 aligned to the genome of *C. elegans*. The draft genome contigs for *H. verbana* were mapped and
163 anchored to the 6 chromosomes of *C. elegans* (Fig. 2A). A majority of the mapped sequences
164 achieved better fit onto chromosomes 1 and X of the *C. elegans* genome (S2 Table).

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

draft genome, which is similar to the first reports of draft genomes for fellow lophotrochozoans *C. teleta*, *H. robusta*, and *L. gigantea,* a gastropod mollusc (Table 2C).

This study is the first to publish an annotated draft genome sequence for the medicinal leech, *H. verbana*. Overall, the genome assembly consists of 250 Mbp and 61,282 contigs, 84.53% of which have been predicted to contain a protein-coding function. The draft genome is also predicted to contain 26,210 protein-coding genes and a repetitive content of 6.67%. The raw short-read sequence data, synthetic long-reads, and assembled contigs for the present study have been deposited into NCBI under BioProject PRJNA551036. The draft genome assembly will assist in providing tools to understand the underlying molecular processes involved in ongoing studies in neurophysiology, developmental biology, and neuroethology that utilize *H. verbana* [63-66]. Whole-genome characterization for Hirudinae *H. medicinalis*, *H. manillensis* [67] and *H. verbana* is expanding and will help distinguish and clarify distinct genetic undertones of these previously amalgamated species. Future efforts to better annotate and complete a genome for *H. verbana* will enable insight into genetic mechanisms of processes investigated with this model organism [66, 68, 69], 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.

Literature Cited

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.

10. Lemke, S., et al., *May salivary gland secretory proteins from hematophagous leeches (Hirudo verbana) reach pharmacologically relevant concentrations in the vertebrate host?* PLoS One, 2013. **8**(9): p. e73809.

11. Nobrega, M.A. and L.A. Pennacchio, *Comparative genomic analysis as a tool for biological discovery.* J Physiol, 2004. **554**(Pt 1): p. 31-9.

12. Clarke, S.L., et al., *Human developmental enhancers conserved between deuterostomes and protostomes.* PLoS Genet, 2012. **8**(8): p. e1002852.

13. Dickinson, D.J. and B. Goldstein, *CRISPR-Based Methods for Caenorhabditis elegans Genome Engineering.* Genetics, 2016. **202**(3): p. 885-901.

14. Gratz, S.J., et al., *CRISPR-Cas9 Genome Editing in Drosophila.* Curr Protoc Mol Biol, 2015. **111**: p. 31 2 1-31 2 20.

15. Simpson, J.H. and L.L. Looger, *Functional Imaging and Optogenetics in Drosophila.* Genetics, 2018. **208**(4): p. 1291-1309.

16. Leifer, A.M., et al., *Optogenetic manipulation of neural activity in freely moving Caenorhabditis elegans.* Nat Methods, 2011. **8**(2): p. 147-52.

17. Kutschera, U. and J. Elliott, *The European medicinal leech Hirudo medicinalis L.: Morphology and occurrence of an endangered species.* Zoosystematics and Evolution, 2014. **90**(2): p. 271-280.

18. Siddall, M.E., et al., *Diverse molecular data demonstrate that commercially available medicinal leeches are not Hirudo medicinalis.* Proc Biol Sci, 2007. **274**(1617): p. 1481-7.

19. Petrauskiené, L., O. Utevska, and S. Utevsky, *Can different species of medicinal leeches (Hirudospp.) interbreed?* Invertebrate Biology, 2009. **128**(4): p. 324-331.

20. Trontelj, P. and S.Y. Utevsky, *Phylogeny and phylogeography of medicinal leeches (genus Hirudo): fast dispersal and shallow genetic structure.* Mol Phylogenet Evol, 2012. **63**(2): p. 475-85.

21. Trontelj, P. and S.Y. Utevsky, *Celebrity with a neglected taxonomy: molecular systematics of the medicinal leech (genus Hirudo).* Mol Phylogenet Evol, 2005. **34**(3): p. 616-24.

22. Bely, A.E. and D.A. Weisblat, *Lessons from leeches: a call for DNA barcoding in the lab.* Evol Dev, 2006. **8**(6): p. 491-501.

23. Macagno, E.R., et al., *Construction of a medicinal leech transcriptome database and its application to the identification of leech homologs of neural and innate immune genes.* BMC Genomics, 2010. **11**: p. 407.

24. Babenko, V.V., et al., *Draft genome sequences of Hirudo medicinalis and salivary transcriptome of three closely related medicinal leeches.* BMC Genomics, 2020. **21**(1): p. 331.

25. Hibsh, D., et al., *De novo transcriptome assembly databases for the central nervous system of the medicinal leech.* Sci Data, 2015. **2**: p. 150015.

26. Northcutt, A.J., et al., *An annotated CNS transcriptome of the medicinal leech, Hirudo verbana: De novo sequencing to characterize genes associated with nervous system activity.* PLoS One, 2018. **13**(7): p. e0201206.

27. Stowasser, A., et al., *Electrophysiology and transcriptomics reveal two photoreceptor classes and complex visual integration in Hirudo verbana.* J Exp Biol, 2019. **222**(Pt 15).

28. Kvist, S., et al., *Draft genome of the European medicinal leech Hirudo medicinalis (Annelida, Clitellata, Hirudiniformes) with emphasis on anticoagulants.* Sci Rep, 2020. **10**(1): p. 9885.

29. Bankevich, A. and P.A. Pevzner, *TruSPAdes: barcode assembly of TruSeq synthetic long reads.* Nat Methods, 2016. **13**(3): p. 248-50.

30. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-5.

31. Tarailo-Graovac, M. and N. Chen, *Using RepeatMasker to identify repetitive elements in genomic sequences.* Curr Protoc Bioinformatics, 2009. **Chapter 4**: p. Unit 4 10.

32. Jurka, J., et al., *Repbase Update, a database of eukaryotic repetitive elements.* Cytogenet Genome Res, 2005. **110**(1-4): p. 462-7.

33. Hubley, R., et al., *The Dfam database of repetitive DNA families.* Nucleic Acids Res, 2016. **44**(D1): p. D81-9.

34. Simao, F.A., et al., *BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.* Bioinformatics, 2015. **31**(19): p. 3210-2.

35. Waterhouse, R.M., et al., *BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics.* Mol Biol Evol, 2018. **35**(3): p. 543-548.

36. Zdobnov, E.M., et al., *OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs.* Nucleic Acids Res, 2017. **45**(D1): p. D744-D749.

37. Stanke, M., A. Tzvetkova, and B. Morgenstern, *AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome.* Genome Biol, 2006. **7 Suppl 1**: p. S11 1-8.

38. Potter, S.C., et al., *HMMER web server: 2018 update.* Nucleic Acids Res, 2018. **46**(W1): p. W200-W204.

39. Finn, R.D., J. Clements, and S.R. Eddy, *HMMER web server: interactive sequence similarity searching.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W29-37.

40. Camacho, C., et al., *BLAST+: architecture and applications.* BMC Bioinformatics, 2009. **10**: p. 421.

41. UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Res, 2019. **47**(D1): p. D506-D515.

42. Gotz, S., et al., *High-throughput functional annotation and data mining with the Blast2GO suite.* Nucleic Acids Res, 2008. **36**(10): p. 3420-35.

43. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-40.

44. Johnson, M., et al., *NCBI BLAST: a better web interface.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W5-9.

45. Holt, C. and M. Yandell, *MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.* BMC Bioinformatics, 2011. **12**: p. 491.

46. Korf, I., *Gene finding in novel genomes.* BMC Bioinformatics, 2004. **5**: p. 59.

47. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.

48. Hillier, L.W., et al., *Genomics in C. elegans: so many genes, such a little worm.* Genome Res, 2005. **15**(12): p. 1651-60.

307 49. Emms, D.M. and S. Kelly, *OrthoFinder: phylogenetic orthology inference for*
308 *comparative genomics.* Genome Biol, 2019. **20**(1): p. 238.
309 50. Buchfink, B., C. Xie, and D.H. Huson, *Fast and sensitive protein alignment using*
310 *DIAMOND.* Nat Methods, 2015. **12**(1): p. 59-60.
311 51. Enright, A.J., S. Van Dongen, and C.A. Ouzounis, *An efficient algorithm for large-scale*
312 *detection of protein families.* Nucleic Acids Res, 2002. **30**(7): p. 1575-84.
313 52. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2--approximately maximum-likelihood*
314 *trees for large alignments.* PLoS One, 2010. **5**(3): p. e9490.
315 53. Robinson, O., D. Dylus, and C. Dessimoz, *Phylo.io: Interactive Viewing and Comparison*
316 *of Large Phylogenetic Trees on the Web.* Mol Biol Evol, 2016. **33**(8): p. 2163-6.
317 54. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a*
318 *single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
319 55. Luo, R., et al., *SOAPdenovo2: an empirically improved memory-efficient short-read de*
320 *novo assembler.* Gigascience, 2012. **1**(1): p. 18.
321 56. Li, D., et al., *MEGAHIT: an ultra-fast single-node solution for large and complex*
322 *metagenomics assembly via succinct de Bruijn graph.* Bioinformatics, 2015. **31**(10): p.
323 1674-6.
324 57. Bankevich, A., et al., *SPAdes: a new genome assembly algorithm and its applications to*
325 *single-cell sequencing.* J Comput Biol, 2012. **19**(5): p. 455-77.
326 58. Boisvert, S., F. Laviolette, and J. Corbeil, *Ray: simultaneous assembly of reads from a*
327 *mix of high-throughput sequencing technologies.* J Comput Biol, 2010. **17**(11): p. 1519-
328 33.

59. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.

60. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastBLAST: homology relationships for millions of proteins.* PLoS One, 2008. **3**(10): p. e3589.

61. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.* Nucleic Acids Res, 2016. **44**(D1): p. D733-45.

62. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.

63. Harley, C.M. and D.A. Wagenaar, *Scanning behavior in the medicinal leech Hirudo verbana.* PLoS One, 2014. **9**(1): p. e86120.

64. Yazdani, N., et al., *Expression of a dominant negative mutant innexin in identified neurons and glial cells reveals selective interactions among gap junctional proteins.* Dev Neurobiol, 2013. **73**(8): p. 571-86.

65. Yuan, S. and B.D. Burrell, *Endocannabinoid-dependent long-term depression in a nociceptive synapse requires coordinated presynaptic and postsynaptic transcription and translation.* J Neurosci, 2013. **33**(10): p. 4349-58.

66. Baranzini, N., et al., *Antimicrobial Role of RNASET2 Protein During Innate Immune Response in the Medicinal Leech Hirudo verbana.* Front Immunol, 2020. **11**: p. 370.

67. Guan, D.L., et al., *Draft Genome of the Asian Buffalo Leech Hirudinaria manillensis.* Front Genet, 2019. **10**: p. 1321.

68. Grey, K.B., B.L. Moss, and B.D. Burrell, *Molecular identification and expression of the NMDA receptor NR1 subunit in the leech.* Invert Neurosci, 2009. **9**(1): p. 11-20.

352 69. Kabeiseman, E., R. Paulsen, and B.D. Burrell, *Characterization of a monoacylglycerol*
353 *lipase in the medicinal leech, Hirudo verbana.* Comp Biochem Physiol B Biochem Mol
354 Biol, 2020. **243-244**: p. 110433.

355

356

Figure Legends

Figure 1 – *H. verbana* draft genome gene ontology distribution for the top 20 most abundant sequence annotations for each classification (biological process, molecular function, and cellular component) at GO level 5.

Figure 2 – (A) Alignment of *H. verbana* draft genome contigs to the chromosomes (I-X) of *C. elegans* (B) Reconstructed phylogenic tree based on orthologous gene families.

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%

Table 1 - (A) Whole genome sequencing reads obtained from Illumina HiSeq for *H. verbana de novo* draft genome assembly (B) Statistics for barcode-assembled synthetic long reads generated using TruSPAdes (C) Summary of assembly statistics of the draft genome for *H. verbana*.

Table 1

Table 2: Analysis of the completeness and orthology of *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%
UniRef90	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

Table 2 - (A) BUSCO statistics assessing the completeness of the *H. verbana* draft genome based on 978 metazoan-conserved genes (B) Summary of structural functional annotation for *H. verbana* draft genome (C) Comparison of *H. verbana draft* genome size and predicted number of genes to 3 closely related spiralian genomes (two annelids and one mollusc).

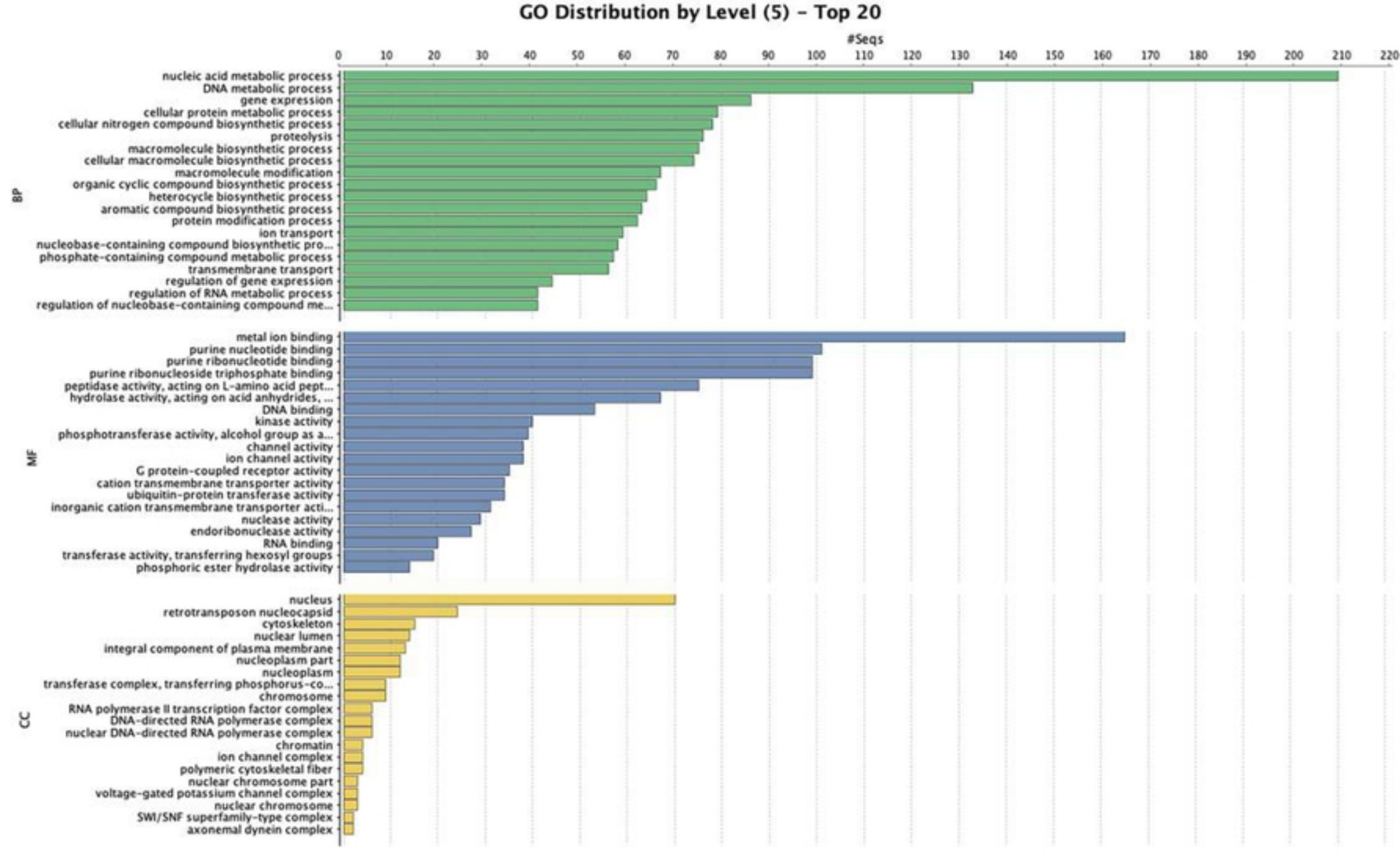
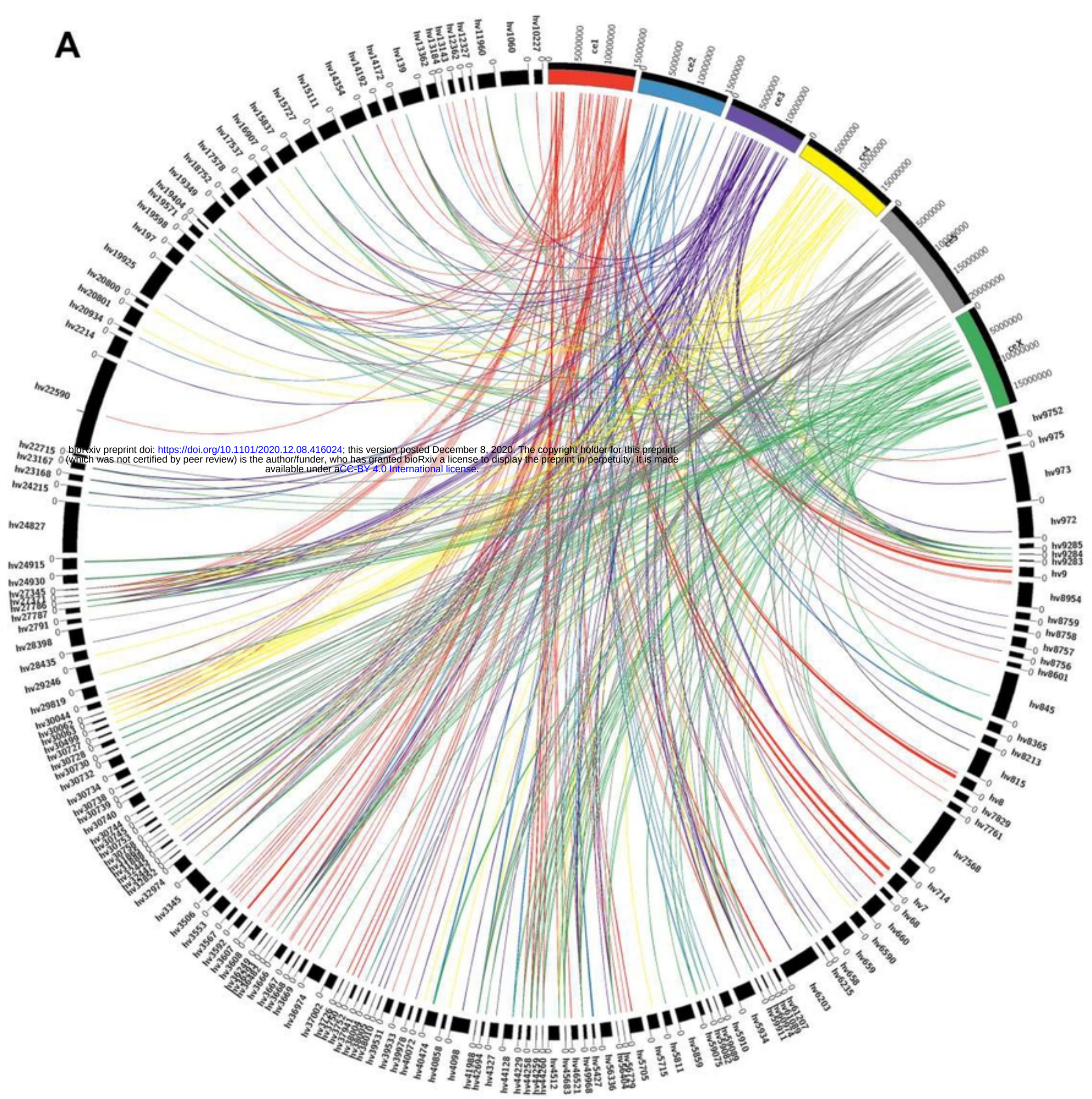

GO Distribution by Level (5) – Top 20
#Seqs
0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220
BP
nucleic acid metabolic process
DNA metabolic process
gene expression
cellular protein metabolic process
cellular nitrogen compound biosynthetic process
proteolysis
macromolecule 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

Figure 1

A



B

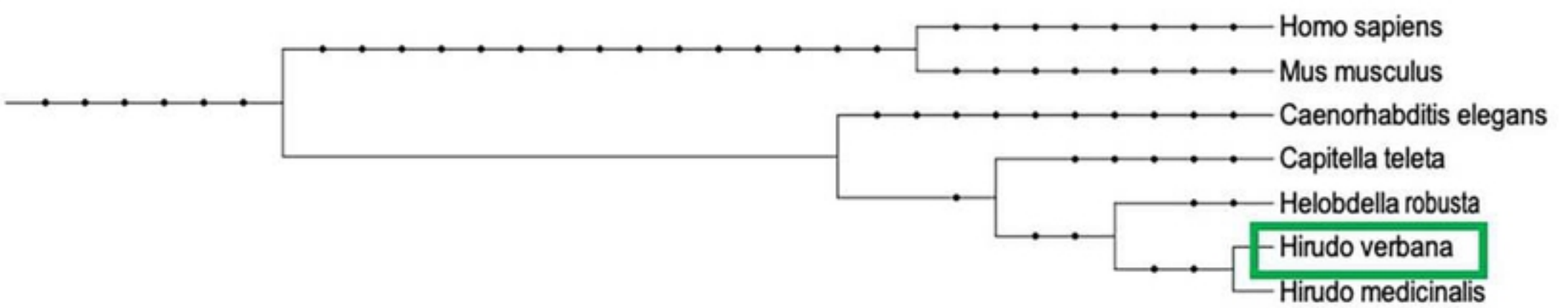


Figure 2